



**U.S. Army Research Institute
for the Behavioral and Social Sciences**

Research Report 1882

**CRITICAL THINKING TRAINING
FOR ARMY OFFICERS
VOLUME TWO:
A MODEL OF CRITICAL THINKING**

Susan C. Fischer and V. Alan Spiker
Anacapa Sciences, Inc.

Sharon L. Riedel
U.S. Army Research Institute

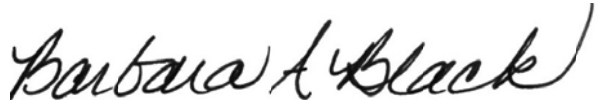
February 2009

Approved for public release; distribution is unlimited.

**U.S. Army Research Institute
for the Behavioral and Social Sciences**

**A Directorate of the Department of the Army
Deputy Chief of Staff, G1**

Authorized and approved for distribution:


BARBARA A. BLACK, Ph.D.
Research Program Manager
Training and Leader Development


MICHELLE SAMS, Ph.D.
Director

Technical review by

Robert Solick, U.S. Army Research Institute
Gregory Ruark, U.S. Army Research Institute

NOTICES

DISTRIBUTION: Primary distribution of this Research Report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPE-ARI-ZXM, 2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

FINAL DISPOSITION: This Research Report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Research Report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE					
1. REPORT DATE (dd-mm-yy) February 2009		2. REPORT TYPE Final		3. DATES COVERED (from... to) January 2004 – November 2006	
4. TITLE AND SUBTITLE Critical Thinking Training for Army Officers Volume Two: A Model of Critical Thinking				5a. CONTRACT OR GRANT NUMBER W74V8H-04-C-0007	
				5b. PROGRAM ELEMENT NUMBER 622785	
6. AUTHOR(S) Susan C. Fischer, V. Alan Spiker (Anacapa Sciences, Inc.), and Sharon L. Riedel (U.S. Army Research Institute)				5c. PROJECT NUMBER A790	
				5d. TASK NUMBER 333	
				5e. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Anacapa Sciences, Inc. U.S. Army Research Institute 301 E. Carrillo St. Fort Leavenworth Research Unit Santa Barbara, CA 93101 851 McClellan Ave Fort Leavenworth, KS 66027				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences 2511 Jefferson Davis Highway Arlington, VA 22202-3926 ATTN: Fort Leavenworth Research Unit				10. MONITOR ACRONYM ARI	
				11. MONITOR REPORT NUMBER Research Report 1882	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.					
13. SUPPLEMENTARY NOTES Subject Matter POC and Contracting Officer's Representative: Dr. Sharon Riedel					
14. ABSTRACT (<i>Maximum 200 words</i>): This report is the second of three volumes describing a multi-year research program to develop and evaluate web based training in critical thinking for Army officers. The first volume presents an overview of the research effort that developed and validated a theoretical model for the training, selected and validated eight high impact critical thinking skills for Army officers, and developed and evaluated the training course. This volume describes the results of a literature review on critical thinking, a model of critical thinking that forms the theoretical basis for the training, and investigations that were conducted to validate the model. Volume Three describes a web-based prototype training system that trains two critical thinking skills. Included in Volume Three are a description of the functional requirements, pedagogical principles, course content, and evaluation of the training. A fourth report (Fischer, Spiker, & Riedel, 2008) describes an expanded version of the training system that provides training for eight critical thinking skills for Army officers.					
15. SUBJECT TERMS Critical thinking, computer-based training, web-based training, critical thinking skill					
SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT Unclassified	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified	Unlimited	102	Diane Hadjiosif Technical Publications Specialist 703/602-8047

Research Report 1882

**CRITICAL THINKING TRAINING
FOR ARMY OFFICERS
VOLUME TWO:
A MODEL OF CRITICAL THINKING**

Susan C. Fischer and V. Alan Spiker
Anacapa Sciences, Inc.

Sharon L. Riedel
U.S. Army Research Institute

Fort Leavenworth Research Unit
Stanley M. Halpin, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

February 2009

Army Project Number
622785.A790

**Personnel, Performance
and Training Technology**

Approved for public release; distribution is unlimited.

ACKNOWLEDGMENTS

The multifaceted research on critical thinking reported in this volume could not have been completed without the contributions of many individuals. We are deeply indebted to each person who has made this work possible. We would like to Dr. Stan Halpin, Chief of the U.S. Army Research Institute (ARI) Fort Leavenworth Research Unit and other members of the ARI staff for their support and review of our research.

Several members of the Anacapa staff were significant contributors to this research. Most notably, Amy Newsam spent many hours searching and reviewing the critical thinking (CT) literature, developing stimulus materials and measures for the validation experiment, interviewing subject matter experts, and any number of assorted tasks that kept the project together. We sincerely appreciate her efforts. Bill Campsey was also instrumental to the success of this project. We are grateful for his support of the Intermediate Level Education (ILE) team in which he helped them use the model of CT described in this report to develop curriculum for officers at the Army Command and General Staff College.

We would also like to thank John Lewis and the ILE team, particularly Jeri Gregory, for their early recognition of the value of our concept of CT. We have learned a great deal about implementing training in CT from them. We are grateful for their hard work, conscientiousness, and patience as they worked to make our ideas a reality in training.

The validation of the model could not have been accomplished without the psychology students who served as participants. We thank them for their service. We are also indebted to Major James Pietsch, who served as an invaluable liaison between Anacapa Sciences and the Army officers we interviewed at Ft. Hood, Texas. We learned a great deal about critical thinking and battle command from the officers of the 1st Cavalry Division. We greatly appreciate their willingness to share their extensive knowledge and insights with us.

Finally, we would like to thank our friend and colleague, Jim Kornell, for his critical review of the model of CT we have developed. His insights have strengthened our work, for which we are sincerely grateful.

CRITICAL THINKING TRAINING FOR ARMY OFFICERS

VOLUME TWO: A MODEL OF CRITICAL THINKING

EXECUTIVE SUMMARY

Research Requirement:

Advanced training in critical thinking (CT) is needed for adult populations in many fields of work. Not surprisingly, the United States military is at the forefront of the effort to promote and improve thinking skills. Military leadership requires the application of high quality CT for effective battle command. As quoted in the Army Field Manual 3-0 (Department of the Army, 2001)

...[It is] essential that all leaders from subaltern to commanding general familiarize themselves with the art of clear, logical thinking. It is more valuable to be able to analyze one battle situation correctly, recognize its decisive elements and devise a simple, workable solution for it, than to memorize all the erudition ever written of war.
(Infantry in Battle, 1939, p. 14)

Clear, logical thinking, or critical thinking, then, is at the heart of leadership. This report is the second of a series of three that describe a research program to systematically and rigorously develop a web based training program in critical thinking for Army officers.

Before effective CT training can be developed, training objectives must be developed and the critical thinking skills (CTS) to train must be identified. Ideally, those objectives and CTS should be derived from an empirically tested theoretical model of CT. However, no such model of critical thinking existed prior to this research. This report discusses the results of a comprehensive review of the literature in critical thinking and model of critical thinking that evolved from this review. This volume also provides the results of experiments to validate the model and to investigate whether the model is applicable to Army battle command. The latter investigation also identified critical thinking skills that were important to and difficult to execute in battle command. The model and identified CTS serve as the basis for the development of a prototype CT training program for Army Officers that is reported in Volume Three of this series.

Procedure:

The primary purpose of this portion of the CT research program was to develop a testable model of CT. In this report, such a model is presented. It is based on current conceptions of CT offered by the philosophy and education literatures and is grounded in current psychological theories of reasoning and judgment.

An experiment was conducted to test some of the central predictions made by the model. Twenty-six participants (5 males and 21 females), ranging in age from 20 to 51 years of age, took part in the experiment. Participants were asked to perform three different tasks related nine

written paragraphs, each of which described a different research investigation. The task instructions asked the participant to either 1) understand, 2) make a judgment about, or 3) simply identify the general topic of the material presented. The type and amount of substantive content of the paragraphs describing the nine research investigations was also varied.

A second investigation was conducted to determine the degree to which the CT model that was developed could be applied to the kinds of problems faced by Army leaders. Eighteen Army officers stationed at Ft. Hood, Texas participated in the investigation. Participants completed a survey that assessed their opinions and experiences concerning CT skills as applied to the domain of battle command.

Findings:

The model that was developed incorporates many ideas about CT offered by leading thinkers in philosophy and education. It embodies many of the CT skills and predisposing attitudes discussed in the CT literature. It also specifies the relationships among a variety of variables that previous researchers have discussed, such as the influence of experience and knowledge, and the relationship of CT to cognitive tasks (e.g., judgment and problem solving). The model, however, goes beyond the largely rational/analytic work conducted to date by providing a framework in which CT can be empirically investigated as a cognitive process.

The validation experiment that tested several predictions of the model yielded a number of results. The results did *not* support the prediction that high substance material increases the tendency to apply CT skills. Under some task conditions, low substance material actually generated *more* CT than high substance material. The results, however, did support the prediction that the application of CT skills is more likely if the available information is degraded in some way, i.e., is conflicting, disordered, uncertain, etc. The prediction that CT is initiated when an individual engages in certain types of tasks was supported. However, it appears that some tasks may not elicit CT equally. These results suggest that refinement of the model may be needed with regard to task. The results of this experiment also failed to support the notion that predisposing individual difference factors affect the tendency to engage in CT skills. The model's prediction concerning experience was supported. It appears that experience does increase the application of CT skills. The findings of the experiment were mixed regarding the hypothesis that CT skills are associated with a corresponding increase in negative affect. Negative affect was not directly related to CT in this experiment.

The results of the second investigation, which tested whether the model could be applied to Army battle command, indicated that the CT model largely captures the critical thinking skills, situational conditions, and predisposing factors important to Army battle command. All of the particular instances of these three variables were regarded as, at least, sometimes important to battle command. Data obtained at Fort Hood confirmed that several battle command tasks are particularly problematic, i.e., officers reported that they had observed deficiencies related to CT in these battle-command tasks. The results of this investigation were used to identify a set of CT skills important and problematic to Army battle command.

Utilization and Dissemination of Findings:

The model of CT generated a number of predictions that previously had not been empirically tested. The model was sufficiently specified to permit falsification of many of its assertions, which other models of CT in the literature had not provided. The present investigation tested five of the model's central predictions. As a result of the investigation we have a clearer picture of the effect of judgment and understanding tasks on CT and the effect of stimulus substance on CT. It is now clear that CT does not always generate negative affect and that experience may well increase CT.

Although the results of the validation experiment were mixed in their support of the model, the model has passed an important scientific criterion. It has generated testable hypotheses that have produced empirical findings from which we have gained knowledge. Some of the findings point to places in the model that require greater specification or modification. Other findings are consistent with the model's predictions.

These results also have practical implications for the design of information systems and for educational and training programs that seek to increase the use of CT skills. Designers and teachers should be aware that people may not question highly substantive material any more than low substantive material. If CT is desired, inconsistent content might be highlighted by information systems. Similarly, if educational and training programs seek to encourage CT, one strategy would be to sensitize students to inconsistent material.

The CT skills identified in this research as problematic to Army battle command may be utilized for training and assessment purposes and to increase self-awareness. These skills have been integrated into the Army Command and General Staff College Intermediate Level Education (ILE) course materials. The original skills identified in the first phase of the present research are now incorporated into each of the major blocks that are being taught in the ILE Common Core course. Also, training concepts discussed in Fischer, Spiker and Riedel (2008b) have been adopted to teach the skills in ILE.

CRITICAL THINKING TRAINING FOR ARMY OFFICERS

VOLUME TWO: A MODEL OF CRITICAL THINKING

CONTENTS

	Page
INTRODUCTION	1
Societal Interest in Critical Thinking	1
Critical Thinking in the Military	2
Purpose of the Research Program	4
REVIEW OF CRITICAL THINKING LITERATURE.....	5
What is CT?	5
What Skills are Involved in CT?.....	14
How is CT Operationalized/Measured?.....	14
What Role do Attitudes Play in CT?.....	19
What Roles do Stimuli and Context Play in CT?	20
Who are the Best Critical Thinkers?	20
Can CT be Trained?	21
What is Missing from the Empirical Research on CT?	22
A MODEL OF CRITICAL THINKING	23
Overview	23
Core Tenets of the CT Model	24
The CT Model.....	25
Contextual Factors	25
Meta-Tasks.....	26
Predisposing Individual Difference Factors.....	27
Controlled Process and Critical Thinking Skills.....	27
Moderating Variables.....	30
Negative Affective Consequences	30
Measures of CT.....	31
VALIDATION OF THE CRITICAL THINKING MODEL.....	33
Method	34
Participants.....	34
Materials	34
Design	36
Procedure	36
Results.....	38
Effect of Task and Stimulus Variables on Indicators of CT	38
Effect of CT on Affect	43
Effect of Moderating Variable (Level of Education) on CT.....	43
Relationship of Predisposing Factor (Need for Cognition) to CT	46

CONTENTS (Continued)

Discussion	48
Effect of Substance of Material on CT	48
Effect of Task on CT	49
Effect of Predisposing Factors on CT	50
Effect of Moderating Variables on CT	50
Effect of CT on Negative Affect.....	51
Summary and Conclusions	52
 AN INVESTIGATION OF CT IN ARMY BATTLE COMMAND.....	53
Method	53
Participants.....	53
Survey Materials	53
Procedure	54
Results.....	55
CT Skills	55
Discussion	59
 SELECTION OF HIGH-PAYOFF CT SKILLS FOR BATTLE COMMAND	60
 CT SKILL TRAINING AT THE COMMAND AND GENERAL STAFF COLLEGE.....	66
Background.....	66
CT Skills Incorporated in Course Curriculum	66
 CONCLUSION.....	68
 REFERENCES	69
 APPENDIX A: CT SKILLS EXTRACTED FROM LITERATURE	A-1
 APPENDIX B: PREDISPOSING FACTORS FOR CRITICAL THINKING	B-1
 APPENDIX C: INSTRUCTIONS USED TO MANIPULATE TASK IN VALIDATION INVESTIGATION	C-1
 APPENDIX D: TWENTY-SEVEN PARAGRAPHS USED IN VALIDATION INVESTIGATION.....	D-3

CONTENTS (Continued)

LIST OF TABLES

Table 1. Sample of CT Definitions, Themes, and Source Disciplines Provided in the Literature	7
2. Number of Propositions per Topic and Substantive Content Type.....	35
3. Sample Descriptive Statistics for NFC and Other Measures	47
4. Rank, Duty Position, and Branch of Participating Officers	54
5. Means and Distribution of Importance Ratings and Frequency of Problems for 13 Broad Classes of CT skills.....	55
6. Distribution of Importance Ratings for 11 Situation Conditions Associated with CT.....	57
7. Distribution of Importance Ratings for 9 Predisposing Attitudes Associated with CT.....	58
8. Core CT Skills Selected for Training Implementation.....	63
9. Relationship of Battle Command Tasks, CT Issues and Selected CTSs.....	64
10. Critical Thinking Skills Incorporated into the Original CGSC-ILE Core Course	67

LIST OF FIGURES

Figure 1. Process model of critical thinking.....	26
2. Mean response time as a function of substance of stimulus material	39
3. Mean reported effort as a function of task and substance of stimulus material	40
4. Mean number of questions of belief as a function of task and substance of stimulus material	42
5. Mean number of questions of belief as a function of substance of stimulus material and educational level	44
6. Mean number of questions of belief as a function of task and educational level.....	44
7a. Mean number of questions of belief asked by undergraduates as a function of task and substance of stimulus material	45
7b. Mean number of questions of belief asked by graduate students as a function of task and substance of stimulus material	45
8. Mean number of checks on thinking as a function of task and education level.....	46
9. Ratings of importance (x) by reported problems (y) for 13 broad classes of CT skills	61

CRITICAL THINKING TRAINING FOR ARMY OFFICERS VOLUME TWO: A MODEL OF CRITICAL THINKING

INTRODUCTION

Societal Interest in Critical Thinking

Interest in promoting critical thinking (CT) skills has increased over the past 20 years in a variety of diverse applications such as public education, military leadership, nursing, technical vocations, and corporate business (e.g., Cohen, Adelman, Tolcott, Bresnick, & Marvin, 1994; Fallesen, Michel, Lussier, Pounds, 1996; Miller & Malcolm, 1990; National Education Goals Panel, 1991; Tucker, 1996). For example, the development of CT skills in students who attend public educational institutions has become a central component of the United States' National Education goals (National Education Goals Panel, 1991). In the military, research on leadership has identified the need for training programs that promote better thinking to improve battle command decision-making, and at least one course has been developed and implemented (Fallesen, Michel, Lussier, & Pounds, 1996). In the public and business sectors, the National League for Nursing has required the demonstration of CT in graduates of nursing education programs (Facione, 1995), DeVry vocational programs now assess the thinking skills of their students (Tucker, 1996), and many US corporations now provide employees with training in thinking skills (Tucker, 1996). In short, increasing CT skills in United States citizens has become a significant goal within government, public, corporate, and military arenas.

Feeding and supporting this increased interest; CT has also become a recognized construct in philosophy, education, and, to a lesser degree, psychology. Unfortunately, authors vary substantially in their stated and operational definitions of CT, and the field is highly fragmented. Yet, despite its messy conceptual state, CT continues to be an issue of concern for a diverse group of interests.

There are at least two reasons why this is so. First, society is experiencing an increased need for intellectual skills due to demands created by developing technologies. In the past 20 years, our economic and social systems have become dependent on complex technologies, with information becoming either a primary or intermediate product that must be processed to serve decision-making. Humans must critically think about and mentally manipulate information to make effective decisions. When information is incomplete, uncertain, or unreliable, the ability to evaluate its quality becomes paramount for competent decision-making. Thus, CT is a skill necessary for the manipulation of information, especially when the information is degraded. In the future, greater reliance on information will necessitate greater reliance on CT skills. All areas of society are experiencing increases in available information and a corresponding greater demand on intellectual processing. In some domains, however, technological development has increased the available information to such a high degree that job demands may soon exceed current skill levels.

Secondly, leaders in several domains (e.g., nursing, business, and military operations) have recognized the need to improve the CT of personnel within their areas (e.g., Hawley, 1998).

This concern is exacerbated by the decline in mid-level positions, with the attendant necessity for information-based decisions to be made by persons having less experience within many organizations. Moreover, it is not surprising that CT skills may be substandard because education in nearly all arenas has traditionally focused on the accumulation of content knowledge, often neglecting to teach the reasoning skills that process such knowledge. Tucker (1996, p.5) remarks that “while content knowledge is a crucial part of this value creation, critical thinking skills are the adjudicative engines that drive everything from boardroom strategic decisions to the creative responses of a software company's help desk.” In short, education and training may not have kept up with changes in skill demand.

Critical Thinking in the Military

It is not surprising that the United States military is at the forefront of the effort to promote and improve thinking skills. Military leadership demands the application of high quality CT for effective battle command, where battle command applies “to the leadership element of combat power...Commanders visualize the operation, describe it in terms of intent and guidance, and direct the actions of subordinates within their intent. They directly influence operations by personal presence, supported by their command and control system” (Department of the Army, 2001, Section 5-1). Military leaders must make tactical decisions in complex and stressful situations where knowledge is incomplete and uncertain. The information provided to a battle commander is always incomplete, often inaccurate, and sometimes purposefully misleading. The use of CT skills to evaluate battlefield information is crucial under these circumstances. Moreover, leaders’ ultimate solutions to battle command problems must be effective, yet not be predictable. Thus, they cannot simply base their battle plans on well-learned battlefield patterns; they must reason through and integrate an enormous amount of information.

As quoted in the Army Field Manual 3-0 (Department of the Army, 2001, Section 1-1),

...[It is] essential that all leaders from subaltern to commanding general familiarize themselves with the art of clear, logical thinking. It is more valuable to be able to analyze one battle situation correctly, recognize its decisive elements and devise a simple, workable solution for it, than to memorize all the erudition ever written of war.

Infantry in Battle. (1939).

Army leaders are also finding themselves in situations that bear little resemblance to conflict situations they have previously experienced or studied. In recent years, they have been expected to serve peacekeeping roles, for example, in which their job is to control conflict among two or more opposing and hostile groups within a foreign country. In such situations, learned rules of engagement and principles of warfare often do not apply. There is no single enemy and battle lines may not exist. In such situations, novel solutions that are the product of CT are likely to be critical to success.

Sometimes, battle commanders must make rapid decisions where there may not be time for extensive CT. Other times, however, a battle commander may have several days to develop his

plan and issue orders, which is sufficient time to apply concentrated thinking skills. In cases where the ensuing battle requires rapid decision-making, even an available 20 minutes may provide an opportunity to apply CT skills. In fact, CT becomes even more important in the execution, as opposed to the planning, phase of battle because events rarely occur in concordance with the original plan. The dynamic nature of battle demands that officers apply high quality reasoning skills to information that may require reassessment of the situation and changes to strategy. CT is not only critical for commanders at all echelon levels, but also for staff officers who are responsible for summarizing large quantities of information from numerous sources, or for making recommendations to the commander. In summary, battle command is clearly a domain in which CT is important to performance.

Extensive training time is devoted to the accumulation of content knowledge in the military. For example, every Army officer studies Clausewitz's principles of warfare. Hours are spent examining and evaluating historical battles to develop a deep understanding of the relationships among the principal factors affecting the outcomes of warfare. Officers and enlisted personnel are also thoroughly trained in the use of available weapons and their capacity. Traditional methods of training tactical decision-making offer a prescriptive model that corresponds to doctrine, and focuses on the products of decision-making (Fallesen, et al., 1996). Historically, less training time has been spent on improving the process of thinking and decision-making¹. It is not that the Army educational system has neglected to provide instruction in critical thinking or reasoning. However, fewer resources have been devoted to the training of thinking processes than to other important skills. Moreover, the prescriptive and procedural nature of the doctrinal methods may actually discourage the application of thinking skills, inhibiting the creation of novel solutions that might be the result of CT (Fallesen, et al., 1996). Army Training Center experiences tend to reinforce the schoolhouse model of doctrinal decision-making by evaluating unit performance based on adherence to a prescriptive model. In doing so, algorithmic approaches to command decision-making are encouraged and dynamic and creative approaches are not reinforced.

If relatively few resources are devoted to developing good thinking habits, officers must develop and hone their own methods of thinking to support decision-making. Without explicit training, whatever thinking skills a military leader possesses are gained through on-the-job experience, fortuitous experiences in training exercises, individual disposition, or other idiosyncratic means such as self-study. Establishing an integrated training program to address the development of thinking skills in battlefield commanders is preferable to hoping that these skills will develop on their own. In short, the education of military commanders seems a prime opportunity for developing training designed to foster thinking skills and that the domain of battle command is one in which CT is crucial. However, the development of effective thinking skills requires extensive deliberate practice and sometimes unpleasant intellectual work (Paul & Elder, 2001). As is necessary to develop any complex skill, developing proficient CT habits also requires excellent coaching, internal motivation, self-awareness, and the ability to critically evaluate one's own performance (Paul & Elder, 2001).

Development of training that effectively and efficiently increases thinking skills paramount to proficient battle command rests on the ability to identify such skills. Before effective training

¹ However, this is changing. For example, the Intermediate Level Education curriculum at the Command General Staff College includes training in critical thinking that is integrated into every unit.

can be developed, training objectives must be specified, and those objectives should be derived from an empirically tested model of CT. However, no such model currently exists. What is needed is a model that provides answers to a series of questions about CT. The obvious first question is “What is CT?” CT must be understood in the context of other known psychological constructs. For example, what is the relationship between intelligence and CT? Is CT the same as reasoning ability? How are other variables (i.e., attitudes, personality, or cognitive traits) related to CT? If CT is a set of thinking skills, as many have suggested (e.g., Paul & Elder, 2001), which skills are included in that set? The extensive literature on CT that has developed since the early 1940s sheds some light on these questions, but also reveals stark gaps in our knowledge. The following section reviews current conceptions of CT based on a review of the relevant literature.

Purpose of Research Program

The primary purpose of this portion of the research program was to develop and validate a testable model of CT. In the third section of this volume, we present such a model, which is based on current conceptions of CT offered by the philosophy and education literatures and grounded in psychological theories of reasoning and judgment. In the fourth section, we discuss a validation experiment that tested some of the model’s assumptions.

A secondary purpose of the present research was to evaluate the role of CT in Army battle command. Military decision-making was selected for investigation because (1) CT is critical to its success and (2) a need for new training systems has been identified that would help military personnel successfully resolve the kinds of missions they now face. An investigation was conducted to determine the degree to which the CT model we developed could be applied to the kinds of problems faced by Army leaders. The investigation demonstrated that a number of CT skills are both important and problematic to battle command. Eight CT skills, in particular, were identified by the investigation. The investigation and CT skills are discussed in sections four and five of this report, respectively.

A subset of the CT skills identified in our investigation of Army battle command were recognized as important to training being conducted at Intermediate the Level Education (ILE) program at the Command and General Staff College at Fort Leavenworth, Kansas. The ILE curriculum developer incorporated the set of the CT skills described in section five of this report into the ILE courseware. The final section of this report discusses how the skills are being used in the ILE curriculum.

REVIEW OF CRITICAL THINKING LITERATURE

Most of what is written about CT can be found in the philosophy and education literatures. To a lesser degree, one may find references to CT as a separate construct in the psychological literature (e.g., Halpern, 1996; Baron & Sternberg, 1986). If one searches existing literature databases using the terms “critical thinking,” one will obtain approximately 600 hits, 400 of which reside in education sources. Most of the rest are located in philosophy books and journals. Thus, it seems fair to conclude that CT is a construct that has been largely developed by philosophers and educators. However, CT was first conceived by two psychologists in the early 1940’s, Goodwin Watson and Edward Glaser. Watson and Glaser (1980) also developed the first test of CT, the Watson-Glaser Critical Thinking Appraisal (WGCTA), which is still the most widely used test of the ability.

Because most of the work on CT has been conducted by educators and philosophers, the construct has not undergone the kind of empirical examination typically conducted by psychologists. Its discriminant validity relative to other, well-established psychological constructs such as IQ, working memory, and reasoning, for example, has rarely been studied. It is the authors’ admittedly subjective opinion that the lack of empirical investigation of CT and its relationship to other individual difference dimensions has produced a fractionated view of the construct. Without the grounding of data, theorists have been free to postulate divergent descriptions of CT. Therefore, the review of CT is begun with the question of its identity.

In the present review of the CT literature, a discussion is also presented of a variety of related issues, including CT’s relationship to other constructs, how it is typically measured, environmental and situational factors that might determine its application, individual differences in CT, whether or not it can be developed with training, and the skills that may underlie CT. The review is concluded with a brief discussion of the most important information missing from the literature.

What is Critical Thinking?

Overview

CT research is not currently guided by a dominant model or theory. In 1990, the American Philosophical Association (APA) used the Delphi technique to develop a consensus definition of CT based on responses of 46 CT experts (APA, 1990). However, the resulting definition (depicted in the first row of Table 1) has not served to focus the CT research. Our review of the literature, which covered the period *since* the APA published its definition, revealed many different conceptions of CT with only a modest degree of overlap. Multiple definitions of a given construct are not rare in the social sciences. The theoretical state in which we find CT is all too common. Consider, for example, the lack of consensus definitions for mental workload and situation awareness. What makes matters worse for CT than for other constructs, however, is that there have been very few studies that seek to tie CT to empirical phenomena or to further its theoretical development based on empirical investigations. Most of the work in this area has been largely analytic, rather than empirical. Lacking hard facts that would set boundaries on CT,

theorists have been free to conceptualize CT in any way that suits their needs, biases, or preferences.

Thus, there is no simple answer to the question of what CT is. In the absence of a dominant theory, or even a consensus definition, it might be prudent to consider the most common measurement instruments of CT to be *de facto* models of CT. The three most commonly used measures of CT are the WGCTA (Watson & Glaser, 1980), the California Critical Thinking Skills Test (CCTST) (Facione, Facione, Blohm, Howard & Giancarlo, 1998), and the Cornell Critical Thinking Test (CCTT) (Ennis, Millman & Tomko, 1985). These tests would at least provide operational definitions of the construct that could be employed across studies to ensure comparability among research efforts. However, the tests' manuals and other studies reveal that each of the three tests suffers from relatively low internal reliability estimates and questionable psychometric markers (Ennis, et al., 1985; Facione, et al., 1998; Jacobs, 1999; Watson & Glaser, 1980). For example, Jacobs (1999) found a lack of comparability and corresponding poor construct validity between the two forms of the CCTST. The inter-correlations among the three tests are also relatively low, suggesting that they tap different abilities. For example, estimates of the relationship between the WGCTA and the CCTST run between $r = .41$ and $.54$ (Facione, et al., 1998). The highest measurements of criterion validity of the CCTST are with the GRE and its subscales, where $r = .719$ for GRE total score, $r = .708$ for GRE analytic, and $r = .716$ for GRE quantitative. Thus, it may be that these instruments are simply tapping variability in performance that could just as well be explained by scores on intelligence and/or achievement tests (Facione, et al., 1998; Tucker, 1996). For these reasons, current tests do not offer an adequate operational, let alone formal, definition of CT. A detailed discussion of these three CT tests is provided later in this report.

Despite differences among conceptions of CT, interpretation of the literature reveals a modest amount of overlap and redundancy. Several "themes" repeat themselves among the definitions. To elucidate the range and variability among CT conceptualizations, as well as emerging themes, a representative sample of 22 definitions is shown in Table 1. The six themes identified in Table 1 are based on our interpretation of each source's definition of CT. Any definition that refers to inference, analysis, reasons, reasonableness, interpretation, or analysis was classified as involving *logic and reasoning*. Many, but not all, theorists regard the ability to use reasoning and informal applied logic as central to CT. Another common theme abstracted from the definitions in Table 1 is the ability to make *judgments*. Definitions or discussions of CT that included reference to evaluation or judgment, usually of a claim, were classified as having a *judgment* theme. Some theorists see judgment as the "critical" component of CT. Others make no mention of judgment in their definitions or discussions. Other themes seem to describe a state of mind rather than a skill or ability. For example, several theorists describe CT as an attitude or activity that is *reflective or questioning*. Definitions that seemed to describe an attitude or statement of mind involving questioning or reflectiveness were classified as incorporating this category or theme. A few theorists describe CT as a recursive, interactive activity that involves *meta-cognition* while others simply note that CT involves some sort of *mental process*. Discussions of CT that referenced self-regulatory, recursive, self-shaping, or meta-cognitive thinking were classified as involving meta-cognition. Those that explicitly discussed CT as a process were classified as having a mental process theme. Finally, several definitions explicitly emphasized the *purposeful* or goal-directed nature of CT.

Table 1. Sample of CT Definitions, Themes, and Source Disciplines in the Literature

Definition of CT	Source	Theme(s)
Purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based	American Philosophical Association, 1990	Judgment, Logic and Reasoning Meta-cognitive Control Purposeful
To be moved by reasons; a willingness, desire, and disposition to base one's actions and beliefs on reasons	Siegel, 1988	Logic and Reasoning Questioning/Reflective State of Mind
The testing and evaluation of proposed solutions provided by creative thinking	Moore, McCann, & McCann, 1985	Judgment
A practical reflective activity that has reasonable belief or action as its goal	Ennis, 1987	Logic and Reasoning Purposeful Questioning/Reflective State of Mind
A systematic, purposeful, disciplined, comprehensive way to form and shape one's thinking that is based on intellectual standards	Paul, 1995	Meta-cognitive Control Purposeful
Art of thinking about your thinking while you are thinking in order to make your thinking clearer, more accurate, or more defensible	Paul, et al., 1990	Meta-cognitive Control
Making reasoned judgments	Beyer, 1995	Judgment Logic and Reasoning
To question definitions, evidence, actions, and beliefs; to consider what is, what was, or might have been, and what may yet be	Moss & Koziol, 1991	Questioning/Reflective State of Mind
The process of evaluating statements, arguments, and experiences	D'Angelo, 1971	Judgment
The careful and deliberate determination of whether to accept, reject, or suspend judgment about a claim	Moore & Parker, 1989	Judgment
Problem solving process that is not restricted by habit or conformity but is free to be uncommon to what seems normal or natural	Perkins, 1993	Mental Process Questioning/Reflective State of Mind
Deciding rationally what to or what not to believe	Norris, 1985	Judgment Logic and Reasoning
Investigation whose purpose is to explore a situation, phenomenon, question, or problem to arrive at a hypothesis or conclusion that integrates all available information and that can therefore be convincingly justified	Kurfiss, 1988	Questioning/Reflective State of Mind
The propensity and skill to engage in an activity with reflective skepticism	McPeck, 1996	Questioning/Reflective State of Mind
Assessment of the logical and empirical adequacy of a nonfictional statement	Baker & Anderson, 1987	Judgment Logic and Reasoning
Involves three elements: (1) an attitude of being disposed to consider thoughtfully, (2) knowledge of logical and reasoning methods, and (3) skill in applying reasoning methods	Glaser, 1941	Logic and Reasoning Questioning/Reflective State of Mind
A style of open-minded reasoning that should be generalizable across content domains	Sa, West, & Stanovich, 1999	Logic and Reasoning Questioning/Reflective State of Mind
Active, purposeful, and organized efforts to make sense of our world by carefully examining our thinking and the thinking of others in order to clarify and improve our understanding	Gadzella & Masten, 1998a	Meta-cognitive Purposeful Questioning/Reflective State of Mind
A cluster of elaborative mental activities involving nuanced judgment and analysis of complex situations according to multiple criteria	Resnick, 1987	Judgment Mental Process
Use of cognitive skills or strategies that increase the probability of a desired outcome. Describes thinking that is purposeful, reasoned, and goal-directed	Halpern, 1997	Logic and Reasoning Purposeful
Sound reasoning, judgment that entails analysis, synthesis, evaluation, and imagination	Dymek, 1999	Logic and Reasoning Judgment
The application of recognition/meta-cognitive processing, particularly different mental models to solve a given tactical problem or make a tactical decision	Cohen, Thompson, Adelman, Bresnick & Riedel, 1999	Meta-cognitive Control Mental Process

There appears to be a tendency for some disciplines to emphasize certain themes more than other disciplines. Definitions that were published in educational and educational psychology documents (e.g., Gadzella & Masten, 1998a; Glaser, 1941; Kurfiss, 1988; Moss & Koziol, 1991; Perkins, 1993; Sa, West, & Stanovich, 1999) tend to describe CT as a questioning or reflective state of mind. In contrast, logic and reasoning seem to be more central to definitions offered by philosophy (e.g., APA, 1990; Beyer, 1995; Siegel, 1988; Ennis, 1987). The idea that CT is a mental process appears more frequently in psychological sources (e.g., Halpern, 1996, 1997; Resnick, 1987; Cohen, et al., 1999). As shown in Table 1, however, these trends exist amidst a large degree of overlap within and among fields. Complicating matters, individual authors sometimes emphasize different CT themes in different documents (e.g., Gadzella & Masten 1998a, 1998b) or even within a single document. For example, Norris (1989) seems to regard Ennis's (1987), Siegel's (1988), and Norris' & Ennis's (1989) definitions as equivalent, when in fact they emphasize different aspects of CT.

It is tempting to describe CT by stringing together the six themes abstracted from the literature review. In other words, one might regard CT as a purposeful mental process that involves logic and reasoning, judgment, and meta-cognition, and incorporates a reflective or questioning state of mind. If so, the amalgam definition would most closely resemble the lengthy one offered by the APA (1990), of which a small part is reproduced in Table 1. However, the APA definition is still missing the mental process theme and under-emphasizes other themes. It is highly unlikely that theorists would unanimously approve the resulting description because controversy still exists on such themes as logic and reasoning (Massaro, 1997).

Considerable disagreement also exists among leading researchers about whether CT is episodic (a state or process limited in time that varies within an individual), dispositional (a tendency to behave in certain ways that varies across individuals), or an ability (a skill that one acquires) (Ennis, 1996; McCarthy, 1996; McPeck, 1996; Norris, 1996; Siegel, 1996). Although the APA (1990) "consensus" view defines CT as being composed of skills and dispositions, it is clear that this definition has not influenced other researchers' perspectives on the topic. Moreover, construing CT as applied logic and reasoning, by far the most common theme emerging from the literature, has come under criticism from "feminists, critical theorists, advocates of hermeneutics, and culture studies" (Massaro, 1997; Walters, 1994). Some researchers regard the emphasis on reasoning as too limiting. They seek to include processes that do not necessarily involve applied informal logic. Their criticism is even greater for the most common operational definitions of CT, which are essentially multiple-choice tests of reasoning ability.

We see no reason why an amalgam definition of these six themes, or any other analytically derived construal of CT, would do any better in providing a definition on which leading theorists could agree. The basic problem in all cases is that CT has not been shown to have strong ties to empirical phenomena (Tucker, 1996). Tucker (1996) and Norris (1996) maintain that a purely rational approach to organizing the construct will ultimately fail. It must be used in conjunction with empirical data. The construct must be shown to have stable and predictable relationships with empirical phenomena or it becomes a concept with little meaning. Norris (1996) notes that

if we want any theory that employs the term critical thinking as a central concept to have anything to do with educating people, then its definition cannot be derived using solely conceptual analysis.” Dominant theories become dominant because they explain empirical phenomena better than alternative theories. Data serves to ground and limit conceptions of constructs. For example, Norris (1996) and Tucker (1996) independently suggest the need to examine the behavior of people regarded as “good critical thinkers” and “bad critical thinkers.” Norris (1996) suggests that an initial focus might be on what, if anything, is common about the thinking of these individuals when they are engaged in thinking in their fields. Tucker (1996) suggests examining cases or instances of CT to discern its properties. We believe there are many ways to model the relationship of CT to empirical phenomena and offer some ideas later in this report.

At a minimum, the CT model should specify the relationship of CT to other cognitive processes that have extensive literature bases, such as pattern recognition, decision-making, thinking, problem solving, situation assessment, reasoning (verbal, inductive, deductive), and creativity. Thus, it would be possible to differentiate among instances of pattern recognition processing, for example, and CT processing. An adequate model would also stimulate further research by generating testable hypotheses regarding the inputs, processes, and products involved in CT.

The Relationship of CT to Other Psychological Constructs

Unfortunately, researchers disagree about CT’s relationships to other psychological constructs. For example, some researchers equate CT with problem solving (Moss & Koziol, 1991; Pellegrino, 1995; Quellmalz, 1987). In particular, Quellmalz (1987) has provided a mapping between CT and problem solving elements, noting their similarities. However, many others see CT as a cognitive activity that merely provides input to other cognitive tasks (Beyer, 1995; Cohen, et al., 1996; Cohen, et al., 1999; D’Angelo, 1971; Halpen, 1992; Lauer, 1998; Swartz, 1998). For example, Lauer (1998) agrees with Gubbins (1986) that CT is a process serving many other cognitive tasks such as problem solving, decision-making, inference making, thinking divergently, evaluating information and sources, and reasoning.

Cohen, et al., (1999) also hold the view that CT serves decision-making. Based on observations of decision-making performance in naval anti-air warfare, Cohen et al.’s (1999) model uses meta-recognition skills to relate CT to decision-making. Cohen et al. (1999) acknowledge the validity of the naturalistic decision-making literature (e.g., Klein, Orasanu, Calderwood, & Zsombok, 1993; Zsombok, 1997), which states that decision-making in natural environments is driven by recognition processes, especially when time is limited and the decision maker’s expertise is high. They also note, however, that recognition processes may not be useful for novel or ambiguous problems. To account for the cognitive events that occur under such situations, Cohen and his colleagues propose a four-step sequence of meta-cognitive processes that serve as a hypothesis testing procedure when information is uncertain. These meta-cognitive processes employ four CT skills: identifying evidence-conclusion relationships, critiquing problems in arguments, correcting problems in arguments, and testing. Thus, for Cohen et al. (1999), CT is used when the preferred recognition processes cannot be applied. In their view, naturalistic decision-making identifies a solution that somewhat fits the problem.

Then CT identifies how the solution is inappropriate, information that is uncertain, and how the original solution should be changed to better fit the current situation.

At least one investigation has shown a positive relationship between CT and creativity (Gadzella & Penland, 1995). CT has also been shown to be modestly correlated with IQ (Glaser, 1941; Royalty, 1995), and with measures of deep processing (Gadzella, Ginther, & Bryant, 1997). Both IQ and CT tests include items that assess reasoning skills, so it is not surprising that a positive relationship has been observed between the two constructs. It is also not surprising that tests of reasoning skills have been shown to be moderately correlated with measures of CT dispositions (McBride & Reed, 1998). For example, one subscale of the Cornell Critical Thinking Test (planning) has also been shown to be correlated with reading (Farley & Elmore, 1992). These findings suggest that part of the variance in CT measures may arise from variation in information processing ability and intellectual power. However, there have been so few studies that this assertion is premature. Additional research is necessary to determine how and why CT is related to other measures of mental processing such as IQ, creativity, deep processing, and reasoning ability.

While the available evidence suggests that CT may have some relationship to mental capacity, it appears to bear no relationship to the content of beliefs. Many theorists have suggested that a marker of poor CT is belief in the paranormal, evidenced by reading one's horoscope on a daily basis, for example (e.g., Halpern, 1998). Glaser (1941) himself thought that measures of CT would be positively related to progressive political attitudes. However, Glaser's findings showed that individual variation in CT is not related to political beliefs. Similarly, others have found that CT is not related to global belief in the paranormal (Royalty, 1995). At this point, it appears that individual differences in CT do not govern patterns of belief.

Psychological research has largely neglected the relationship between CT and other forms of thinking. In fact, the psychological literature contains few references to CT and when it does reference CT, the references tend to be found in educational psychology journals and books (e.g., Gadzella & Masten, 1998a, 1998b; Resnick, 1987; Sa, West, & Stanovich, 1999). Separate from a relationship to CT, however, there is considerable research on thinking and the related concepts of judgment and reasoning.² Current theories and empirical research are relevant to CT, although there has been little effort to establish any correspondence. For this reason, a brief overview of the long history of this research is provided here.

Overview of Thinking, Judgment, and Reasoning Research

Prior to the early 1970's, the dominant psychological model based judgment on the calculation of the probability and utility of various options. Proponents of the model thought that judgment was a *rational* choice that utilized a logical reasoning process. Although the rational choice model took on a variety of forms, all versions posited a rational actor who made appropriate calculations of probability and/or utility, and selected the option that had the highest value. However, even as early as the 1950's, researchers began to notice discrepancies between

² There is much overlap in the empirical bases and theoretical accounts of the thinking, judgment, and reasoning research. For this reason, we treat them as a unitary literature.

the predictions of the classic model of judgment and actual behavior (Meehl, 1954; Simon, 1957).

In the early 1970s, an alternative theory proposed that people use heuristics, as opposed to the rational weighing of relevant factors, to make judgments. The theory was, and continues to be, supported by evidence that shows that people make systematic errors of reasoning that are the result of heuristic processing. It is important to understand that the classic rational choice model of reasoning predicts that mistakes will be made but that the pattern of mistakes will be random. Thus, the Tversky and Kahneman (1971) seminal research that demonstrated systematic errors of reasoning provided falsifying evidence of the classic model.

The pattern of errors observed in the 1971 paper and many others (e.g., Tversky & Kahneman, 1973, 1974, 1981, 1982, 1983) was consistent with the idea that people use heuristics to make judgments. An example of a heuristic-based judgment is the now famous case of “Linda”, originally documented by Tversky and Kahneman (1983), but replicated by others (Stanovich & West, 2000). When participants read a paragraph about Linda (in italics below), the vast majority fall prey to an error Tversky and Kahneman (1983) called the conjunction fallacy. Most people (somewhere around 80%) rank the first of the statements that follow the paragraph below as more probable than the second. Of course, the first statement cannot be more probable because it subsumes the second. The probability of the first statement is ranked higher than the second because it is more similar to the given description of Linda. Studies have shown that even statistically-savvy people use a similarity heuristic, rather than their knowledge of probability, to make the judgment.

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

1. *Linda is a bank teller and is active in the feminist movement.*
2. *Linda is a bank teller.*

Since Tversky and Kahneman’s early studies, a great deal of empirical research has been conducted that has repeatedly demonstrated the limits of human reasoning as demonstrated by systematic errors. The cumulative data show a pattern that could not be produced by rational decision makers (e.g., Evans, 1989; Kahneman, Slovic, & Tversky, 1982; see Kahneman, 2003). Despite controversy about the source of the limitation (Stanovich & West, 2000), it is eminently clear that peoples’ reasoning performance is not consistent with the classic or normative model that dominated theories prior to Tversky and Kahneman (1971).

The main point of the research on heuristic judgments was not to demonstrate how irrational people can be (Kahneman, 2000). Much has been learned about the types of heuristics people frequently use (e.g., availability, representativeness, anchoring and adjustment), but that was also not an original goal of the research. Kahneman (2000) states that the central objective of the research was to develop a better understanding of the psychology of intuitive judgment and choice. Tversky and Kahneman did not propose that all judgments were made intuitively, just that there was a tendency to use intuitive processes to make many judgments. They posited that judgments are made from two different systems (Kahneman, 2003). One they labeled

intuition, which they regard as quick, automatic, and implicit. It uses associational strengths to arrive at solutions. The other they called simply, reasoning, which they considered to be effortful and deliberately controlled.

Tversky and Kahneman's theory is only one of many dual process theories that have been developed to explain empirical data on reasoning. Several researchers have recently noted the striking similarity among at least twelve dual-process theories of reasoning (Holyoak & Spellman, 1993; Kahneman, 2003; Sloman, 2002; Stanovich & West, 2000). Stanovich and West (2000) offer a prototype summary of these theories. Each theory posits one process that is quick, effortless, automatic, holistic, implicit, and uses associational strengths as its computational basis. Each also describes a second process that is slow, effortful, deliberate, analytic, explicit, and is based on a rule-based computational system. Stanovich and West (2000) refer to the implicit associational type of process as *System 1* and the conscious deliberate process as *System 2*. Holyoak and Spellman (1993) also distinguish the two processes based on their computational architectures: one uses a parallel connectionist structure while the other uses symbolic serial production systems.

Dual-process theories of judgment have been developed within a variety of psychology sub-disciplines. Most have arisen from the investigation of cognition. Johnson-Laird's (1983) investigation of reasoning and inference-making with mental models, Schneider and Shiffrin's (1977) work on visual search, Posner and Snyder's investigation of attention (1975), and Sloman's (1996) and Evans' (1989) reasoning research are all examples of theories that have been developed within the cognitive psychology tradition. The dual process distinction has been made in applied psychology as well, notably in human factors. Klein (1999) and Rasmussen, Pejtersen and Goodstein (1992) have attempted to explain patterns of judgment found in naturalistic environments. In particular, Klein's observations of countless real-world experts have lead to his control theory, namely that powerful naturalistic decision-making proceeds on the basis of an automatic recognition-based processing akin to System 1. Rasmussen, et al. (1992), adds a third type of processing to his control theory by distinguishing between conscious problem solving activities and the application of simple rules. With this distinction, he divides the prototypical System 2 processing into two types. Within social psychology, the investigation of persuasion and attitude development has also produced several dual-process models (Chaiken, Liberman, & Eagly, 1989; Petty & Cacioppo, 1986). These models attempt to explain how people judge the validity of persuasive messages and acquire prejudice.

Dual-process theories typically posit that judgments can be made using either system (Chaiken & Trope, 1999; Kahneman, 2003; Klein, 1999; Sloman, 1996). Some models posit that System 1 is the preferred system because it demands fewer resources and is less effortful to apply (Chaiken, 1980; Klein, 1999; Rasmussen, et al., 1992). Others, however, believe that the two processes overlap in their domains and are highly interactive (e.g., Kahneman, 2003; Sloman, 2002). Most posit that Systems 1 and 2 run in parallel and that one function of the controlled deliberate process is to monitor the products of the automatic process. The empirical evidence supports the idea that people use associational methods, as well as controlled reasoning, to make judgments. However, under certain characteristic situations the associational processes will systematically produce errors. Kahneman (2003) suggests that the analytic mode of System 2 is primarily used to endorse, make adjustments to, correct, or block the judgment of System 1.

However, if no intuitive response is accessible, System 2 may be the primary processing system used to arrive at judgment (Kahneman, 2003). Sloman (2002) states that the systems work hand in hand as “two experts who are working cooperatively to compute sensible answers.”

Relationship of CT to Dual-Process Theories of Thinking, Reasoning and Judgment

No one, to our knowledge, has attempted to apply current dual-processing theories of reasoning to CT. However, the similarities between descriptions of CT and descriptions of System 2 processing are striking. The words “effortful, controlled, deliberate, purposeful, and conscious” are frequently used for both. If Kahneman (2003) and others are correct in their proposal that errors of reasoning are often made because System 1 is biased toward accessible information based on associational strengths, it may be that CT skills, which are processed by System 2, have not been applied to the situation. In short, reasoning may break down when System 2 has failed to serve its monitoring and correcting function. The application of a dual-process theory may serve to explain many problems of CT noted in the literature. For example, dual-process theories may explain why people seem reluctant to engage in CT, why we see so many errors in reasoning, why people are influenced by superficial characteristics of reasoning problems, and why people hold irrational beliefs (Halpern, 1996; Paul & Elder, 2001).

The Purpose of CT

Examination of CT’s purpose may help to clarify its meaning. We have already discussed the assertion made by several theorists that CT serves higher-level purposeful tasks such as problem solving and decision-making. Moore, McCann, and McCann (1985) posit that CT serves decision-making because it allows the evaluation of sources. Ennis (1987) emphasizes how the judgment element of CT leads to better decision-making because it allows determination of appropriate actions and beliefs. In the applied setting of military decision-making, Cohen, et al. (1999) have asserted that CT allows Army officers to form models of their own actions in regard to enemy decision-making, and then use the models to develop proactive, predictive, and reaction plans. Cohen, et al., (1996) believe that CT improves the accuracy of an officer’s assessment of battlefield situations.

Many philosophers contend that CT’s role in decision-making has the added and lofty benefit of improving society. CT has a long history of being regarded as critical to societal health because it is tantamount to rationality and good judgment. In his seminal work, Glaser (1941) linked CT to a Jeffersonian model of society, stating that good citizenship depends on the ability to critically think. He also suggested that CT benefits society by encouraging cooperation among its members. Lauer (1998) asserts that CT brings sanity to human affairs. D’Angelo’s (1971) view of the purpose of CT is consistent with Lauer’s and Glaser’s. He asserts that CT eliminates undesirable attitudes, beliefs, biases, and prejudices. Also, Beyer (1995) states that CT is important because it is used to recognize and produce things of value, which is critical to societal health.

CT may have ramifications beyond instilling the good judgment that creates societal cooperation. Several researchers hold that CT is responsible for many cultural developments. Holyoak and Spellman (1993) asserted that explicit thinking, perhaps a close relative of CT,

allows us to create new thought, and to adapt the environment to ourselves rather than be adapted to the environment. They note that explicit thinking lets us imagine what is not the case, and what might be. If this is true, and if explicit thinking corresponds to CT, then CT may be responsible for new philosophical and scientific advancements. A similar case is made for intelligence, which some consider to have a close correspondence with CT. Sternberg (1987) states that intelligence allows us to function effectively in complex and changing contexts. Paul (1995) also suggests that CT affords adaptation to new situations. Halpern (1997) offers the idea that CT benefits the individual as well as the group, letting us determine the probable cause of an event and deal with uncertainty.

In summary, CT is commonly regarded as purposeful and is often posited to serve other cognitive functions such as decision-making, problem solving, and judgment. It may also benefit societal health and generate advancements in philosophy and science.

What Skills are Involved in CT?

CT might be better understood by examining the kinds of skills that philosophers believe are integral to CT. Skills that have behavioral outcomes can ground CT in observable performance. Philosophers have proposed a great many skills associated with CT. However, it is not surprising that one's theoretical definition of CT strongly influences the CT skills that he/she identifies. For example, Moore's and Parker's (1989) and Beyer's (1995) emphasis on judgment in their definitions of CT is reflected in their lists of skills, which primarily involve the evaluation of arguments, claims, explanations, and reasoning. In contrast, Moss and Koziol (1991) operationally define CT skills that involve the unbiased extraction of information from text, which is consistent with their conception of CT as a dynamic process of questioning and reasoning. Theorists who emphasize reasoning and logic in their definition of CT tend to include skills such as forming and testing hypotheses, evaluating syllogisms, and avoiding logical fallacies in their stated lists of CT skills (e.g., Ennis, 1987; Glaser, 1941; Siegel, 1988). Thus, we see a strong correspondence between definitions of CT and the skills identified by CT theorists.

The literature is replete with discussions of a large number of CT skills. Appendix A lists 120 CT skills extracted from the literature after literal redundancies were removed. However, conceptual redundancies are still present in the list. For example, skill #28, "recognize fallibility of own opinion," is nearly identical to skill #29, "recognize the probability of bias in your opinion." The skills have been categorized into four types: interpreting, reasoning, assessing, and monitoring.

How is CT Operationalized/Measured?

How CT is measured may also clarify the concept of critical thinking. The need for measuring individual differences in CT has driven development of assessment instruments. Some researchers have developed their own measures of CT as demanded by the particular needs of the investigation. However, the measurement of CT is dominated by three assessment instruments. All three of the tests, which are reviewed below, assess the ability to draw valid conclusions and judge the validity of inferences. As noted previously, logic-based conceptions

of CT have come under criticism from a variety of groups who have a different perspective on CT. Some critics seek to add intuition and creative elements to the analytic skills that largely dominate the concept of CT. Others want to include empathy, imagination, and patterns of discovery to balance the emphasis on objectivity (Walters, 1994). The three tests reviewed below, although the most commonly used, represent a limited approach to CT.

Tests of Critical Thinking

The three most widely used tests of CT are the Watson-Glaser Critical Thinking Appraisal (WGCTA), the Cornell Critical Thinking Test (CCTT), and the California Critical Thinking Skills Test (CCTST). Although each test has unique advantages and disadvantages, they all share some critical features. Specifically, all are multiple-choice tests, come with packaged summaries and test specimens, and have been administered to a number of normative groups throughout the country. The following paragraphs briefly summarize the main features of each test, including relevant psychometric data.

Watson-Glaser Critical Thinking Appraisal (WGCTA). The WGCTA is the oldest test of CT, developed in the early 1940s by Glaser for his doctoral dissertation (Glaser, 1941). The original test has been revised several times, most recently in 1980 (Watson & Glaser, 1980). The authors of the WGCTA define CT as a composite of knowledge, skills, and attitudes. However, the WGCTA measures only skills. For this reason, the WGCTA is not a comprehensive test of CT as it is defined by Watson and Glaser (1980). The WGCTA is an 80-item multiple-choice test, composed of five 16-item subtests. The subtests cover: inference (discriminating among degrees of validity), recognition of assumptions, deduction, interpretation (weighing evidence to determine if certain conclusions are warranted), and evaluation of arguments (distinguishing among strong/weak and relevant/irrelevant arguments).

The WGCTA comes in two equivalent, alternate Forms, A and B. In each form, the first subtest, Inference, is a series of five-foil items where groups of 5 to 6 statements refer to a descriptive paragraph. For each statement, the examinee must decide whether it is (a) true, (b) partly true, (c) there is insufficient data to answer, (d) partly false, or (e) false. The remaining four subtests consist of two-choice items. For the Recognition of Assumptions subtest, examinees must determine whether the assumption in the descriptive text was either made or not made. In the Deduction subtest, examinees must gauge whether conclusions follow or do not follow from the explanatory material. Similarly, the Interpretation subtest asks respondents to indicate whether the conclusion follows or does not follow from the given assumptions. Finally, the Evaluation of Argument subtest asks that examinees note whether a given argument is strong or weak.

Psychometric analyses of WGCTA have yielded mixed results. Because the test has been explicitly designed to measure five separable components of CT, tests of internal consistency are computed separately for each of the five. Split-half reliability coefficients have varied considerably, depending on the population taking the test. Among high school and nursing students, the coefficients range from .69 to .76, somewhat low for a test of this type. On the other hand, reliability coefficients are consistently in the low to middle .80's for populations of college and graduate students.

Validation studies of the WGCTA have looked at its correlations with other tests that have known validity. Correlations vary widely with the population tested. For example, the WGCTA exhibited correlations of .60 and .41, respectively, with the verbal and math Scholastic Aptitude Test (SAT) of freshmen at a large Northeastern university. Conversely, the test is less related to verbal and math SAT scores when administered to nurses ($r = .45$ to $.48$). The Inference and Interpretation subscales have been shown to be significantly correlated with tests of creativity (Gadzella & Penland, 1995). In another investigation, the WGCTA correlated significantly with tests of Deep Processing ($r = .35$) and Fact Retention ($r = .18$) (Gadzella & Masten, 1998b). In addition, the WGCTA exhibits moderate correlations with several other standardized tests, including the Miller Analogies test ($r = .55$), the Wechsler Adult Intelligence Scale ($r = .55$), and the College Entrance Examination Board ($r = .54$, verbal; $r = .43$, math).

The WGCTA possesses a number of positive attributes that make it a useful research tool. However, it measures only a narrow slice of the composite skills of CT. Its main weakness is that it does not comprehensively measure CT. First, it does not assess knowledge and attitudes, two of the components that define CT according to Watson and Glaser. Second, the WGCTA provides a good assessment of analytic skills such as deduction, but there is little in the test that measures induction or generalization.

Cornell Critical Thinking Test (CCTT). Developed in the 1960's, the CCTT comes in two forms: Level X for grades 4-14 and Level Z for advanced and gifted high school students, college students, and other adults (Ennis, Millman, & Tomko, 1985). Our review will focus on the Level Z test because it is most pertinent to adult populations.

The test is based on Robert Ennis' (Ennis, 1962) conception of CT as a general form of cognitive processing that can be dissected and subcategorized into three types of inferences (induction, deduction, evaluation) and four types of bases for those inferences. Bases include the results of inferences, observations, statements made by others, and assumptions. Thus, Ennis argues that a general test of CT must include, besides inference, the credibility of evidence, identification of assumptions, and determination of meaning. Like the WGCTA, the CCTT does not attempt to measure attitudes or predispositions to CT.

The composition of the CCTT's factor structure is less orthogonal than the WGCTA, as the former's subtests comprise overlapping items (Ennis, Millman, & Tomko, 1985). The Level Z test is divided into four parts, consisting of 52 three-foil multiple-choice items. Part I focuses on having examinees judge whether a fact supports a stated hypothesis. In Part II, examinees are given passages of text and asked to judge the credibility of observation reports. In Part III, they must decide which of the listed alternatives would be true if the information given is true. Finally, Part IV asks examinees to form judgments about what is being assumed in an argument.

The Level Z test has been normed on an impressive number of diverse groups, including undergraduates at various universities and graduate students from different fields. As expected, mean scores on the test increase across grade level, with graduate students having the highest scores (Ennis et al., 1985). Tests of reliability, as measured by the Kuder-Richardson split-half coefficient, vary from .50 to .77. The low values are indicative, in part, of the heterogeneous

nature of the items, suggesting to the test authors that the statistics “underestimate” the extent to which comparable test scores would be repeated across multiple administrations (Ennis et al., 1985). Unfortunately, reliability estimates made through pooling split-half coefficients within each subtest were not reported.

Because the target population for the Level Z test is smaller, there has been fewer validation studies of it compared to the Level X test. There have been, however, some interesting uses of the test. For example, scores on Level Z correlated between .2 to .4 with graduate school success, achieving comparable predictability to the Graduate Record Exam (GRE) and the Miller Analogies Test. The test’s authors report that scores on Level Z do not correlate highly with personality variables, except for a negative correlation with dogmatism. Interestingly, the CCTT correlates higher with the math section of the SAT (.51) than the verbal section (.36), a pattern opposite to that found for the WGCTA. The two tests themselves correlate quite highly, .79, when administered to both undergraduates and graduate students.

Within the literature, the WGCTA enjoys more psychometric support than the CCTT, owing to the former’s availability in equivalent forms, data concerning technical quality, and a more extensive normative database (Jacobs, 1999). In a detailed investigation of the CCTT structure, Frisby (1992) concluded that the Level Z test needs the addition of some easier items in order to more accurately distinguish between low- and medium-ability students.

The sections in the CCTT concerning deduction, credibility of evidence, identification of assumptions, and determination of meaning are very clearly written and comprehensive. However, in our opinion, the two parts of the test that address induction are flawed. In each section, a long series of questions is presented, each to be considered independently, where test takers must assume the truth of various assumptions and determine whether the presented information supports the conclusion, refutes the conclusion, or does neither. However, the sequence of different assumptions may produce carryover effects on latter questions. Second, several questions present information concerning replication of the described experiment under new settings (e.g., different countries, variety of ducks). Specification of the correct answer in each case requires that the examinee make some assumptions about which conditions of experimental replication are appropriate, assumptions which may be outside the domain of CT. Third, several questions contain information that is supportive (or refutive) of the conclusion and irrelevant. In cases where there is a mix of relevant and irrelevant information, it is incumbent upon the examinee to subjectively weight the relative importance of the information in selecting his/her response. Such judgments add to the “noise” in the test data and degrade the reliability of the instrument.

California Critical Thinking Skills Test (CCTST). The newest of the three CT tests reviewed here, the CCTST was developed in the early 1990’s by Peter Facione and is based on the APA’s Delphi Project definition of CT. From that meeting, participants concluded that the core cognitive skills of CT are interpretation, analysis, inference, evaluation, and explanation. Thus, a good critical thinker is one who gives “reasoned consideration to evidence, context, theories, methods and criteria in order to form this purposeful judgment” (Facione, Facione, Blohm, Howard, & Giancarlo, 1998). Along with the skills mentioned above, the authors of the CCTST added a sixth, self-regulation, to encompass the meta-cognitive aspects of reasoning.

The CCTST was originally provided in two equivalent, alternate Forms, A and B. The authors published a third version, the CCTST 2000, which is designed to tap a wider audience (beyond college) by including graphics and more varied sample materials. Each test consists of 34 four-foil multiple-choice items. Because the test is short, its factors are not orthogonal as each item contributes to multiple factors (i.e., inference, interpretation, etc.). Also, it is impossible to evaluate the CCTST because answer keys are not made available by the publisher. Unlike the publishers of the other tests, California Academic Press has elected to retain proprietary rights over the answers and factor structure in order to provide services for scoring and administration of tests.

Because CCTST 2000 is relatively new, the psychometric data have been exclusively collected on Forms A and B. Normative data have been collected on three large samples, including 781 college students, 153 nursing students working toward their masters, and 224 enrollees in a law enforcement academy. The mean score for the nurses was significantly higher (19.0) than for the college (15.9) and law enforcement students (14.6). This ordering is consistent with the other CT tests where scores are seen to increase with education level.

From a psychometric standpoint, the CCTST has received mixed marks. Its short length was designed to support administration during a 50-minute college class period. Because of the short length, it isn't surprising that reports of test reliability have proven disappointing, with Jacobs (1999) noting that Forms A and B produced reliability coefficients of .57 and .61, respectively. These statistics were obtained from sample sizes in excess of 700, making their quantitative estimates quite stable. Moreover, Jacobs (1999) has also found that the difficulty levels between the two forms are different, refuting the publisher's claim of "equivalence" between the two tests.

Assessments of criterion validity have shown the CCTST to be significantly correlated with a number of other instruments. These include the verbal ($r = .72$) and math ($r = .58$) components of the GRE, the verbal ($r = .55$) and math ($r = .44$) components of the SAT, and the WGCTA ($r = .41$ to $.54$). However, the inability to decompose the test into stable factors has limited researchers' ability to look for more fine-grained relationships with other test instruments.

Conclusions

The three dominant operational measures of CT reflect the multidimensional nature of the construct. Each test purports to assess subscales that represent separate CT skills discussed in the literature (e.g., inference, evaluation of arguments, detection of unstated assumptions). Thus, reliability estimates of each test suffer because the items tap fundamentally different skills. Unfortunately, reliability estimates for the subscales of the tests either have not been assessed or have been relatively low (Facione, et al., 1998; Jacobs, 1999; Watson & Glaser, 1980). Jacobs' (1999) analysis of the CCTST (Forms A and B) revealed problems with particular items as evidenced by low item-test correlations. Thus, low reliabilities may also be due to poor item design for one or more of the tests. Whether low reliability estimates are due to item problems or the multidimensional nature of the tests, it is clear that current measures of CT fall short of the

psychometric standards typically applied to psychological assessment instruments. Yet, there is no fundamental reason why the assessment of CT should be psychometrically problematic. We suspect that these tests simply have not been scrutinized by researchers nor refined to the degree that other instruments (e.g., the SAT, GRE, Wechsler, and Stanford-Binet) have been. Despite the psychometric issues, we see that, by design and by analysis, the tests tap a loosely structured multidimensional construct. Current conceptions and operational definitions of CT do not describe a single process or ability, but rather, a collection of processes or abilities.

In conclusion, the available tests of CT do not provide as clear picture of CT as one would want. They appear to be measures of applied informal logic. As noted previously, this limited operational and formal definition of CT has come under criticism by other researchers who seek to broaden the concept's meaning. They are also limited in that they do not measure one component of CT that most theorists posit, which is a dispositional attitude that favors its use. In the next section of this review, the role attitudes play in CT is discussed.

What Role do Attitudes Play in CT?

The role of dispositional attitudes in CT has been acknowledged since Glaser's (1941) seminal work, in which he proposed that one essential component of CT was "an attitude of being disposed to consider thoughtfully" (Glaser, 1941, p. 10). Since then, the idea that attitudes influence CT has found its way into many other theorists' writings, including the APA's consensus definition (e.g., APA, 1990; Beyer, 1995; Ennis, 1987; Kurfiss, 1988; Paul, 1995; Perkins, Jay, & Tishman, 1993; Resnick, 1987; Walsh & Hardy, 1999). Many authors regard dispositions as an integral element in CT, equal in status to CT skills. Facione, Facione, and Sanchez (1994) state that mere training to use CT skills does not necessarily create good critical thinkers. Paul and Elder (2001) are explicit about the attitudes they believe are important to CT, which they call characteristics of mind. These include intellectual integrity, humility, sense of justice, perseverance, fair-mindedness, confidence in reason, courage, empathy, and autonomy. A list of attitudes extracted from the literature and hypothesized as important to CT can be seen in Appendix B. Yet, none of the major assessment instruments evaluate dispositions despite their acknowledged influence on CT and their inclusion in formal definitions of CT.

There is, however, one test that purports to measure attitudes, the California Critical Thinking Disposition Inventory (CCTDI) (Walsh and Hardy (1997). Consistent with the APA's formal definition of CT, this instrument is designed to assess seven distinct sub-dispositions: truth-seeking, open-mindedness, analyticity, systematicity, confidence, inquisitiveness, and maturity. However, very little research has been conducted on this test. It is relatively new and is not used as frequently as the WGCTA, CCTST, and CCTT (Walsh & Hardy, 1999). One exception is an investigation of differences among university majors and the sexes on the CCTDI (Walsh & Hardy, 1999). Their results showed that English, psychology and nursing majors scored higher on the CCTDI than history, education and business majors. Also, females scored higher than males on the overall test, and on its subscales of open-mindedness and maturity. Walsh and Hardy (1997) also investigated the factor structure of the CCTDI and showed that the loadings were cross correlated across subscales. Thus, some of the subscales required renaming to reflect these factor loadings. Walsh and Hardy (1997) extracted only two factors that were

stable across men and women. Factor I was identified as perspicacity/confidence and Factor II as receptivity/open-mindedness.

In summary, many authors propose various dispositions or attitudes as being necessary to CT. However, the only instrument designed to assess predisposing attitudes appears to assess only two sources of attitudinal variability, not the many sources it is designed to test. Thus, our knowledge about the influence of dispositions and attitudes on CT is limited.

What Roles do Stimuli and Context Play in CT?

The CT literature contains almost no reference to stimuli or contextual influences on CT. The research and theoretical positions on CT almost uniformly treat CT as a trait, disposition, or state of mind. Thus, even when CT is considered as a mental or meta-cognitive process, researchers have not taken the next step, which is to investigate stimulus, situational, or environmental variables that might influence the quality of, entry into, or maintenance of CT. It is reasonable to expect that stimulus and contextual variables affect the likelihood one will engage in critical thinking and the quality of one's CT. From a psychologist's perspective, the failure to consider stimulus and contextual variables is an enormous gap in the literature. From other researchers' perspectives (Norris, 1996; Tucker, 1996), the most important research issue that needs to be addressed is individual differences in CT ability. These authors reason that, if we cannot distinguish among individuals who critically think well from those who perform poorly, the concept of CT has little practical merit.

Who are the Best Critical Thinkers?

Are some groups of individuals better at CT than others? Investigation of the question, "*Who are the best critical thinkers,*" may ultimately lead to a better understanding of the underlying processes and skills involved in CT. Thus, studies that compare extant groups, based on subject variables, on their demonstrated CT abilities are important to the development of CT as a psychological construct.

CT ability appears to develop according to maturational processes. Researchers have shown that CT improves as children age from grades 4 through 12 (Frisby, 1991). Frisby found that the developmental trend is nonlinear, as the biggest change appears when children graduate from junior to senior high school. Another investigation showed that college seniors do better on tests of CT than college freshman (Keeley, 1992). However, additional research in this area is needed to assess developmental changes in adulthood. More investigation is needed to determine whether CT ability, relative to one's peer group, changes in adulthood.

Certain areas of academic specialization seem to be associated with higher CT ability levels. For example, several researchers have found that students who study psychology, natural science, and English score higher on measures of CT than do students in other majors, particularly those in sociology, social work, nursing and criminal justice (Gadzella & Masten, 1998b; Lawson, 1999). English majors also have been shown to demonstrate attitudes deemed important to CT, such as truth-seeking, open mindedness, confidence, inquisitiveness, and

maturity (Walsh & Hardy, 1999). In contrast, athletes have been found to be less open minded, inquisitive, and mature, independent of their sex (McBride & Reed, 1998).

In general, college students perform poorly on measures of some CT skills. In particular, they are very poor at identifying important unstated assumptions in text (Keeley, 1992). When asked to find assumptions, their most common response is to restate a premise or identify an insignificant assumption. However, a significant amount of variation exists among college students. High achieving students, as measured by GPA, for example, tend to do better on CT tests than low achieving students (McCutcheon, Apperson, Hanson, & Wynn, 1992; Royalty, 1994). Royalty (1994) found that measures of achievement such as GPA and class standing were modestly correlated with CT, as measured by the CCTT. Another investigation found that high achievers tend to make fewer errors on the WGCTA (McCutcheon, Apperson, Hanson & Wynn, 1992). Gadzella, Ginther, & Bryant (1997) found that *A* students produced higher scores on the inference, deductions, and interpretations subscales of the WGCTA than did *C* students.

It is not clear if or how gender is related to CT ability. At least one investigation has shown that females score higher on certain dispositional variables theoretically linked to CT (Walsh & Hardy, 1999). However, many more have found no gender differences with regard to CT.

In summary, too little research has been conducted relating subject variables to CT abilities to determine the individual factors that produce, or are associated with, good critical thinking. The existing research indicates that CT increases with age and experience, but little is known about changes in adulthood. Some college disciplines are associated with high levels of CT, but it is not clear whether the area of investigation encourages development of CT or whether individuals with high levels of CT are drawn to such disciplines. It also appears that high levels of achievement are associated with high CT abilities. But, again, causal relationships cannot be established at this point in time. Moreover, we do not know the relative contributions of raw ability versus training on CT performance.

Can CT be Trained?

Glaser (1941) conducted the seminal work showing that training programs have a positive effect on a variety of CT variables. Glaser developed and administered a program of instruction for students in four high schools. The training covered a variety of topics, including interpretation of text, assessment of underlying assumptions, and other topics of applied logic. Students who received the training did much better on subsequent tests of CT than did students not in the program. Glaser's finding is typical of many other studies that have assessed a variety of CT programs. For example, McBride, and Bonnette (1995) found that CT can be fostered in at-risk groups through training and education.

The issue of training CT has also been examined for particular applied fields, such as military leadership and nursing. Youssef and Goodrich (1996) evaluated the effects of nursing education on development of CT and found that current nursing programs do little to increase the use of CT in their students. However, they also found that CT performance predicted who would pass the nursing licensing exam, suggesting that nursing training could do more to prepare

students for their professional work. For the Air Force, Schaub (1991) found that students can be trained toward a disposition that facilitates CT. Moreover, Schaub showed that this disposition is retained after training. Similarly, Cohen et al. (1996) found that their training program increased Army officer performance on a number of CT variables. For example, trained officers who read arguments in text generated more disconfirming statements, more supporting statements, and provided problem assessments that were similar to subject matter experts'. They also had more significant insights about solving a military scenario and more officers changed their planned courses of action based on the training. In another investigation conducted by Cohen and his colleagues, students identified more gaps in assumptions and information after training (Cohen, Thompson, Adelman, Bresnic, & Riedel, 1999).

The existing research on CT training is encouraging in that it suggests that CT ability and disposition are sensitive to intervention. The relative contributions of innate abilities related to CT should be investigated, however. It may be that CT ability can be generally improved in most people, but that some may be more affected by training and experience than others. This remains an important empirical question.³

What is Missing from the Empirical Research on CT?

The body of research reviewed in this report has not adequately dealt with a number of important issues relevant to CT. For example, few studies have investigated the underlying cognitive processes that produce individual variation in CT. The work relating CT to other psychological constructs and abilities is a good start in clarifying CT's position relative to other human information processing but many questions remain. Perhaps the most significant gap in the existing body of research is the lack of studies that investigate the effects of stimulus and contextual variables on CT. The tendency to think about CT as a trait or disposition may have kept researchers from investigating this basic issue.

Further, little research has been devoted to the effects of CT on human performance in real-world tasks. An exception is Gonzalez's (1996) investigation that showed individual differences in CT may be related to nursing performance. In this investigation, nurses who scored higher on CT measures produced more accurate diagnoses.

What is most needed is a model of CT that is sufficiently specific to be subjected to empirical scrutiny. Such a model could be used to organize and test many of the hypotheses contained in the literature. A model would also help to develop operational definitions and would point to future areas of research. An empirically-grounded model is also essential to any effort seeking a meaningful and significant increase in CT skills because it would identify those skills to be trained. A good model would explain the relationship between CT and other cognitive processes (e.g., decision-making, problem-solving, pattern recognition), identify the relationships between the CT skills and real world tasks, and develop training requirements and curricula that would foster better thinking. The next section describes a model of CT that is grounded in the literature on CT and addresses the requirements cited here.

³ A variety of pedagogical techniques and methods intended to increase levels of CT have been used and tested. However, these techniques bear greater relevance to the topics in the third volume of this report. Thus, they are not discussed here.

A MODEL OF CRITICAL THINKING

A model was developed that incorporates many ideas about CT offered by leading thinkers in philosophy and education. It embodies many of the CT skills and predisposing attitudes discussed in the CT literature. It also specifies the relationships among a variety of variables that previous researchers have discussed, such as the influence of experience and knowledge, and the relationship of CT to cognitive tasks (e.g., judgment and problem solving).

The model, however, goes beyond the largely rational/analytic work conducted to date by providing a framework in which CT can be empirically investigated as a cognitive process. Up to this point, CT has been treated as an individual difference variable that can be measured and correlated with other variables of interest. We offer a model that makes specific testable predictions about the factors that influence CT and about its psychological consequences. The model organizes the most important factors, in the authors' judgment, discussed in the large body of literature on CT. It also establishes the relationships between CT and cognitive processes posited by current theories of cognition and performance. However, its greatest contribution is that it provides a direction for further research. The following discussion of the model begins with an overview of its main features.

Overview

Generally stated, CT is a cognitive process that intervenes between a set of initiating situational conditions and the observable performance of one or more tasks. It is purposeful and deliberate cognitive processing that includes checks on the process and products of thinking and serves other higher-level tasks such as decision-making. CT involves the application of least one of a particular set of cognitive skills that demand the use of meta-cognitive and/or recursive control, consciously controlled logic, or thorough examination of a problem. The execution of CT skills is powered by an effortful, yet flexible, computational process that is capable of controlled meta-cognitive and recursive monitoring of thinking. The processing engine is distinguished from the quick, yet powerful, recognition-based processing (Klein, 1998) that depends on association strengths and that is subject to well-documented accessibility errors (Kahneman, 2003). The application of CT skills has accompanying affective consequences, which may be measured using standard physiological preparations or psychometric instruments.

CT is considered in this model to occupy a relatively brief time frame (in the range of 5 to 30 minutes) in which certain essential processing functions are executed, depending on the desired goal. One can, however, string together a series of CT episodes held together via meta-cognitive monitoring processes, which would define a much longer CT period that might encompass a training exercise, a class, or some important real-world event (e.g., an entire intelligence briefing). If a given individual shows a tendency to initiate CT skills on a frequent basis, he or she might be labeled as a "critical thinker." For a variety of reasons, individuals also vary in the quality of their critical thinking and the quality of the measurable products of the application of CT skills. Critical thinking skills are considered both measurable and modifiable via experience. Therefore, it should be possible to develop and evaluate a robust training program to promote CT skills. Furthermore, if one can identify particular skills or classes of

skills that are more commonly associated with high-performance outcomes, then a major effort to train those skills should yield high-payoffs for participating organizations.

Core Tenets of the CT Model

The CT model rests on five fundamental assumptions or core tenets. First, the CT model assumes that CT is intimately linked to a set of initiating contextual conditions that increase the likelihood that an individual will execute a CT skill. In this way, the model is intentionally distanced from a large segment of the educational and philosophical literature, which has focused on identifying essential CT dispositions and other subject traits for the expressed purpose of improving the educational system, making better citizens, or some other general societal goals. While these larger societal end-points are of value, the approach embodied in the CT model is geared toward concentrated investigation of CT as a psychological, context-governed phenomenon. Specifically, the focus is on identifying the processes that are brought to action in response to definable initiating conditions. The objective of the CT model is a compact yet powerful theory of the skills required to deeply process informational material in response to measurable initiating conditions. Thus, our orientation is toward *a priori* predictions of why a CT skill has been executed, i.e., the contextual conditions that were present at the time it occurred, rather than a non-empirical, *post hoc* examination of possible traits the subject may have brought with him/her to the situation. One area for investigation is determining the types of conditions that promote the use of CT skills.

Second, CT is assumed to be a highly stimulus-bound phenomenon in which effortful, controlled cognitive processing is imposed on a circumscribed stimulus-information complex to achieve some objective. Here the stimulus is defined as being able to be physically isolated and specified in advance. Thus, stimulus materials might either be brief text (e.g. one paragraph), a graphic, some spoken discourse, or a scene from the real world. For battle command applications, the stimulus might be an explanatory paragraph from an operations order (OPORD), a terrain map, a situation report, or the concluding observation in a verbal intelligence briefing. This stands in contrast to traditional CT approaches in which the stimulus materials are either extensive (e.g., an entire class, a book, a newspaper article) or perhaps unspecified, where the objective is to engender improved CT in the person as a whole. The specification of CT as a stimulus-bound phenomenon is unique to our theory. However, it is essential to any psychological theory of CT that seeks to investigate external factors that govern the use of CT skills. This theoretical position also generates a host of testable hypotheses regarding stimulus conditions and their effects on CT.

Third, CT is assumed to be time-limited, where an individual may execute CT skills for only a few minutes (or less). Traditionally, the literature has been vague on the subject of how long people engage in CT. Most theorists seem to consider CT to be fairly extended in time, lasting at least the time required to listen to a class lecture or read a newspaper article. However, a more likely reality is that the application of CT skills is continually being interrupted in order to consult with others, view the consequences in the external world, or simply “take a break” from the mental effort. An area for investigation will be to identify the stimulus and contextual conditions that precipitate these breaks and to determine if countermeasures can be developed to increase persistence in the application of CT skills.

Fourth, CT is assumed to have consequences in which the individuals experience emotions, motivations, and other phenomenological experiences that are reportable. While the literature has made passing reference to CT as being “effortful” or “work,” we propose that its consequences are a vital aspect of CT as a phenomenon. In particular, negative experiential consequences of CT would offer a plausible explanation as to why people do not engage in it more often.⁴ Besides the negative affect generated during CT, there may be other properties of CT that might have implications for research. For example, it may be that the extended application of CT skills produces a state much like immersion does in virtual reality environments, in which people are not easily distracted. An inability to be distracted might then serve as a marker of a CT episode. Similarly, the ability to report the application of CT skills provides experimenters with an independent means to assess whether their attempts at manipulating the stimulus conditions were successful.

Finally, a logical conclusion derived from the previous four tenets is that CT can be subjected to experimental manipulation given proper controls. This is perhaps the most important assumption, as it implies that one can manipulate the likelihood of CT through systematic changes in the stimulus and/or contextual conditions. This view stands in sharp contrast to the bulk of the literature that treats CT as a response measure and an indicator of individual differences, in which people’s scores on a standardized test of CT are correlated with other subject characteristics. The CT model posits that through judicious manipulation of context and stimulus dimensions, the tendency to engage in CT processing will be altered. For example, this suggests that intentionally embedding logically inconsistent phrases in a sentence might encourage subjects to use their interpretative and reasoning skills to process the entire paragraph, thereby increasing the likelihood that they apply CT.

The CT Model

Figure 1 depicts the CT model. It is placed in the context of cognitive processes (i.e. System 1 thinking) posited by the most current models of reasoning and judgment. The model incorporates six major factors involved in CT, including (1) situational conditions, (2) meta-tasks, (3) CT skills, (4) predisposing individual differences, (5) moderating variables, and (6) negative experiential consequences. Each of the major factors is described below.

Contextual Factors

The opportunities for the application of CT skills are set in motion by contextual factors, which include (1) situational conditions and (2) meta-tasks. Situational conditions are environmental factors such as stimuli and surrounding contextual variables. As shown in Figure 1, two types of situational conditions are distinguished, defining and predictive. Both types are posited to influence the use of, and serve as input factors to, CT skills.

⁴ In the interest of parsimony, all individuals engaged in CT are assumed to experience negative consequences such as fatigue, stress, etc. However, further testing may reveal individual differences in the affective experience evoked by CT, which would require refinement of this assumption.

Defining situational conditions are necessary for the application of CT skills. The model posits that two conditions must be present in the situation. These include that (1) the stimulus material to be processed contains substantive information, and (2) time is not severely limited, i.e., sufficient time is available to engage in an appropriate CT skill. Predictive situational conditions are not necessary for the execution of a CT skill, but they increase the likelihood that an individual will engage in CT. Conditions include stimulus variables such as the presence of conflicting information, disordered or unorganized material, uncertain information, and complex material. For example, situations in which conflicting information is presented would tend to, but would not necessarily, elicit CT. Predictive situational conditions also include variation in the problem and the valued outcome of the situation. If there is no single answer to a problem or the outcome has high stakes, the likelihood that one will use CT is increased.

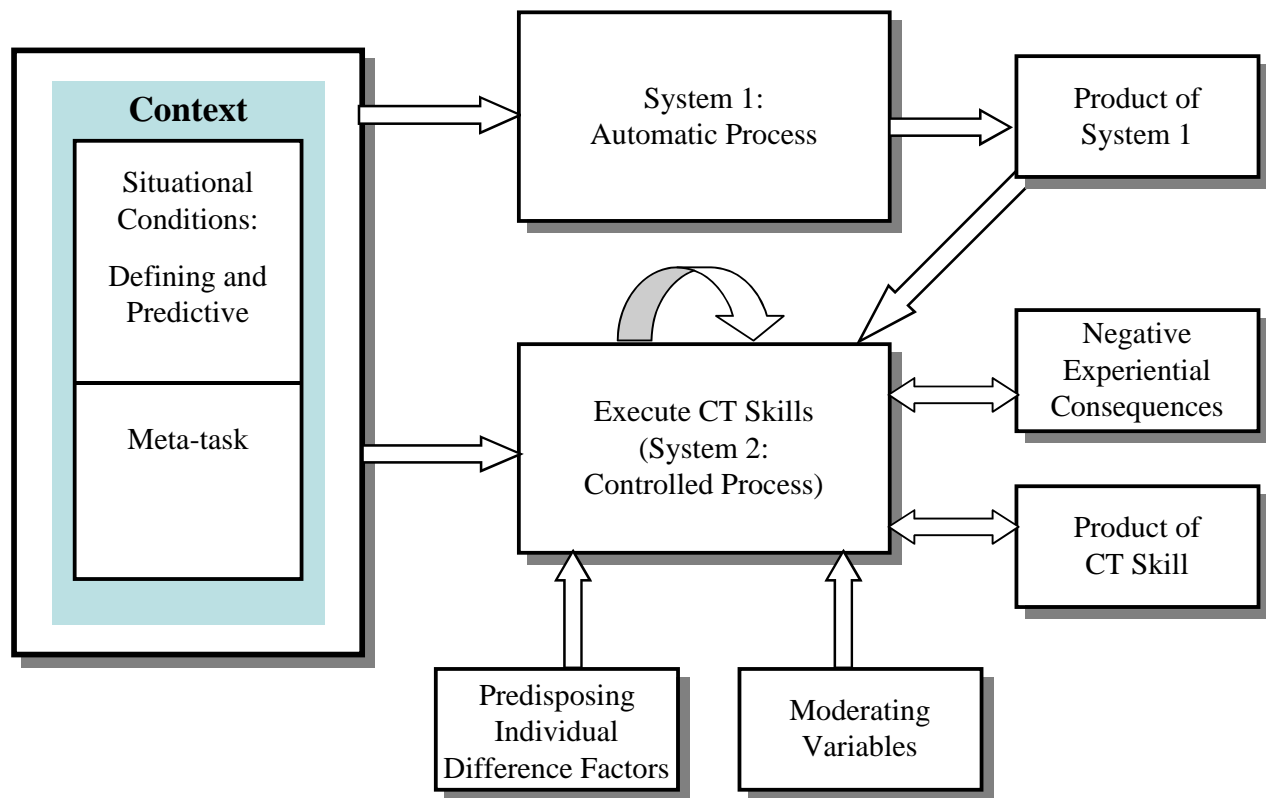


Figure 1. Process model of critical thinking.

Meta-Tasks

Meta-tasks are necessary for and define the general purpose of the application of CT skills. The inclusion of meta-tasks in the model is consistent with definitions of CT that emphasize the process's purposeful nature. The CT model posits that an individual must have as his/her objective to accomplish at least one of four meta-tasks to engage in a CT skill. These include the requirement to understand, make a judgment, make a decision, or solve a complex problem such as creating a plan. The interaction between CT and the meta-tasks posited by the CT model is

important because it begins to establish CT's relationship to other well-researched cognitive processes such as problem solving and decision-making. CT is not an end in itself but serves other objectives specified by the meta-tasks. The meta-tasks also dictate the specific response that will be required to successfully end the execution of a CT skill.

Predisposing Individual Difference Factors

In addition to contextual factors, predisposing individual difference variables also make an individual more or less likely to engage in CT. They influence the likelihood of a person using, or persisting in using, CT, and like the situational context, they serve as input conditions. They do not, however, include individual differences that represent the quality of thinking. Appendix B contains a list of predisposing factors posited to influence CT. Predispositions vary in their strength of relationship to CT. Some may be key factors that strongly affect an individual's use of CT. Other factors may have a weaker relationship to CT, perhaps increasing the likelihood of engaging in CT by a marginal amount. In summary, predispositions are *measurable subject variables*, whether *fixed* or *modifiable*, that influence *use* or *persistence of use* of CT.

The factors listed in Appendix B were selected from many that may be found in the literature. Predisposing factors whose definitions make them inherently immeasurable have been intentionally omitted. Also excluded are those that are simply a restatement of a desired CT skill. For example, we do not include "predisposed to draw general inferences from specific facts". Rather, we consider only predispositions that could be measured with a pretest and should be related to an individual's *likelihood* of using CT skills when faced with the opportunity to do so. We have, in this way, attempted to avoid the circular reasoning trap that ensnares many emerging psychological theories (Meehl, 1950).

Controlled Process and Critical Thinking Skills

The CT model posits that a controlled process (labeled as System 2 in Figure 1)⁵ is used to execute critical thinking skills. If the predisposing factors, meta-tasks, and situational conditions are sufficient, CT skills will be executed by System 2. It is important to note that we do *not* claim that all processing executed in System 2 involves CT skills, but simply that all critical thinking requires the kind of analytic, rational processing that System 1 cannot provide. For example, System 2 may drive other deliberate processes such as controlled visualization used in planning. However, we do not consider controlled visualization to be a CT skill.

Figure 1 also depicts the relationship between contextual factors and System 1 or automatic processing. Contextual factors serve as input to an automatic process (labeled as System 1 in Figure 1), as well as to a controlled process that executes CT skills (System 2). We posit that the information provided by the context feeds into the automatic process as well as the consciously controlled process that drives CT. The automatic processing system is rapid, relatively effortless, and can be performed concurrently with other tasks. In contrast, we posit that critical thinking utilizes System 2 as its processing engine, which is characterized by slow, controlled, serial processing that is effortful, rule-governed yet flexible in its application. In

⁵ The use of the terms "System 1" and "System 2" is based on Stanovich's & West's (2000) terminology for these processes.

Kahneman's terms (2003), System 2 is the process that is used when one engages in reasoning, as opposed to intuition.

Consistent with several theoretical views (Evans & Over, 1996; Kahneman, 2003; Kahneman & Frederick, 2002; Sloman, 2002; Stanovich & West, 2000), both processes are thought to run in parallel and interact as tasks are executed. Specifically, the association-based products of System 1 provide input to System 2, which is used to execute CT skills. System 1 is truly an automatic and uncontrolled process. Therefore, it cannot be initiated or stopped. For this reason, System 2 monitors only the products, but not the process, of System 1. Because System 1 is quick, it often arrives at a solution before System 2. However, System 2 may override the conclusions of the automatic process. Therefore, System 2 has the potential for controlling performance, although it may not always utilize that potential.

System 2 is required for executing CT skills⁶, as we define them, for several reasons. First, a meta-cognitive component that monitors and checks the quality and products of thinking is integral to CT skills. According to most dual-process theories, the quick associative and heuristic processing of System 1 is not designed to handle the recursive nature of meta-cognition. Thus, skills that involve checks on thinking clearly fall in the domain of System 2. Kahneman (2003) posits that one of the central functions of System 2 is to monitor the rapid judgments that System 1 produces, clearly a meta-cognitive task. Therefore, our position (i.e., skills that require meta-cognition [e.g., CT skills] employ System 2 processing) is consistent with Kahneman's (2003) view.

Second, CT skills, as described by many philosophers and educators (e.g., Paul & Elder, 2002), require complete examination and thorough processing of the problem at hand. For example, one may seek a clear statement of a problem by careful and thorough examination of all relevant problem elements. System 1 is not designed for slow, methodical, processing of stimulus content. Automatic processing uses quick recognition of salient and accessible features of available information. In contrast, System 2 does provide the kind of controlled, meta-cognitive checks that are necessary to ensure that all components of a situation are given due consideration.

Third, many CT skills also involve controlled checks on the process and products of logical reasoning. While it is possible to reason deductively and inductively using the quick recognition-based processing of System 1, the resulting conclusions are sometimes prone to error. System 1 typically derives only one solution (Kahneman, 2003). It works to narrow possible action paths, which can be highly effective when the task must be accomplished quickly and the problem space is limited. However, when the problem space is novel or complex, or when solutions must be innovative, System 1 processing can lead to a failure to consider multiple potential solutions. In such situations, these associative processes can lead to a tendency to "jump to conclusions". In contrast, the application of CT skills driven by the System 2 process can be used to logically derive multiple solutions that can be examined and evaluated using the same controlled logic reasoning. We conclude that tasks that involve the use of controlled reasoning and logic can be reliably executed only by using System 2.

⁶ A refined list of CT skills offered by leading philosophers and researchers is given in Appendix B.

Fourth, CT skills typically place heavy demands on cognitive capacity. Critical thinking often involves manipulation of a large number of variables to produce novel solutions, which demands the slower, but more flexible, System 2 processing. In other words, the cognitive demands of the task sometimes exceed the processing capabilities of System 1.

In summary, System 2 cognitive processing is required for the execution of CT skills because (1) meta-cognitive checks are involved in the skill, (2) the skill involves thorough processing, (3) the skill involves controlled reasoning, or (4) the problem to be solved is sufficiently novel or complex such that System 1 cannot produce an acceptable solution. Therefore, System 2 must be used to perform CT skills.

System 2 also provides the meta-cognitive capabilities to monitor the progress of its own processing. In this sense, System 2 has recursive properties, as represented by the curved arrow leading out and back into the “Execute CT Skills” function in Figure 1. We posit that System 2 processing incorporates a monitoring process that determines when the meta-task (judgment, understanding, etc.) has been completed. Thus, successful completion of the meta-tasks as determined by System 2 can also provide input that terminates CT. The self-monitoring contained within System 2 serves to either sustain or terminate the CT episode.

It is important to clarify the functions of Systems 1 and 2 and their relation to CT. First, the *task* posed by a particular situation should not be confused with the particular cognitive *skill* applied to achieve that task. For example, one may have the task of understanding a message, e.g., a commander’s intent statement that could be solved using associational processes of System 1 or controlled CT skills powered by System 2. Therefore, an individual who is trying to understand an intent statement may, or may not, be using a CT skill to do so. An understanding of an intent statement may be produced by either System 1 associational processes or by System 2 controlled processes that involve the application of CT skills.

An even more important point to make is that we do not claim that the application of CT skills driven by System 2 always produces the best solution to a task. In fact, Klein (2002) and many others have shown that expert performance, which can be reasonably used as a standard, is often based on associational processing. However, expert performance is not faultless, and many experiments have shown that even experts make systematic errors in judgment (Kahneman, 2002). While System 1 is usually highly effective at producing quick and powerful solutions, it does not always do so. The associational processes of System 1 that make expert performance so powerful are the same processes responsible for the systematic errors that have been observed under certain conditions (e.g., Tversky & Kahneman, 1973, 1974, 1981, 1982, 1983). While those associational processes become more accurate and refined with experience, they may also lead to a tendency to come to unchecked conclusions and to inappropriately narrow one’s focus. In conclusion, the quality of the solution may be superior if it is derived from the application of a CT skill, it may be superior if derived from System 1 associational processes, or the quality of the two processes’ products may be indistinguishable. Therefore, it would be a mistake to encourage exclusive use of System 2 controlled thinking because that strategy would deny the power and effectiveness of System 1. Similarly, it is not advisable to only use associational processes because controlled deliberate reasoning can (1) produce superior solutions and (2)

provide necessary checks on the products of System 1. Moreover, the issue of which system is most effective is practically irrelevant because most theorists believe that both are almost always used in conjunction to produce a solution. How well the two systems interact probably contributes as much, or more, to the quality of observable and measurable performance (e.g., decision-making or judgment) as the effectiveness of each system alone.

The quality of a solution produced by the application of a CT skill may also be affected by how well the skill is executed. Decrements in performance may be produced by failing to apply a component of the CT skill (e.g., failing to clarify ambiguous information in a message or failing to consider alternative explanations for a pattern of data), failing to accurately perform a component of the skill, or by lacking sufficient knowledge that can be processed by the CT skill. Therefore, one could apply a CT skill and still produce inferior solutions to a task. Moreover, it is not possible to determine whether System 1 or System 2 was applied to derive a solution based on the solution alone. The quality of a solution may also be affected by moderating variables such as educational level and experience. We discuss these in the next section.

Moderating Variables

The model also includes a set of moderating variables representing individual difference factors that, while not directly or causally related to the use of CT skills, mitigate the performance of CT skills. Moderating variables are not factors that represent the stimulus situation. Instead, they reflect individual characteristics (e.g., domain expertise, recent experience, education) that influence how, and how well, CT skills are performed. The moderating factors are also distinct from predisposing factors in that they do not influence whether one executes a CT skill. For example, the resulting products of the CT of a highly educated individual will most likely be superior to those produced by a less educated individual. Thus, there should be a correlation between CT performance and education level. However, high levels of education do not necessarily serve as a gate into CT.

Negative Affective Consequences

An interesting corollary to the present approach is the hypothesis that individuals who engage in CT for any substantive length of time experience affective reactions that, by and large, are not positive. Consistent with the CT literature, we hypothesize that individuals who sustain CT do so in part because they are able to maintain a neutral emotional state. In other words, we propose that there are no positive intrinsic rewards associated with the application of CT skills, although there are probably positive outcomes. Rather, we posit that CT produces largely negative affective consequences. Several affective reactions are specified, the predominant byproduct being one of mental fatigue and increased effort associated with heavy workload. Other affective consequences of CT include increased anxiety, cognitive dissonance, and social awkwardness. In Figure 1, affective consequences are depicted as a byproduct of CT. Negative experiential consequences also serve as input to the decision to maintain a CT episode, as depicted in Figure 1 by the bidirectional arrow. We posit that when the affective consequences of applying a CT skill become too negative, the motivation to maintain the episode is decreased. If the negative consequences are sufficiently strong, they may result in a cessation of the episode.

The negative affective consequences work in conjunction with the controlled process of System 2 to create a CT ‘state’, which may be measurable with physiological or psychological indicators. In other words, it may be possible to independently measure whether someone is applying CT skills, which could produce converging validation of this aspect of the model. The CT model provides a framework for investigation of non-intrusive methods (e.g., time estimation technique used to measure mental workload) as well as psychometric methods to derive independent confirmations of a CT episode.

Measures of CT

We hypothesize that individuals vary in their performance of CT in three ways, each of which may be measured. First, some individuals are more likely to use CT skills when the context calls for it, or even when it does not. While the automatic processing of System 1 will always be used simply because of its associational computational foundation, some individuals may exhibit a greater frequency of employing the more controlled, flexible process of System 2 to meet the demands of meta-tasks than other people. Secondly, individuals vary in their ability to apply CT skills. For example, some may be better able to “identify assumptions in an argument” than others. This is most likely the source of variability that determines performance on common tests of CT, logic, and reasoning. It is also where knowledge, experience, and training are most likely to have their positive effects on measurable performance. Thus, training might have as its objective to improve the ability to execute CT skills. A third source of variability is the ability to apply the products of the appropriate CT skills proficiently to the given meta-task. For example, an individual may have successfully executed a CT skill, but cannot apply the results of his/her analysis to the task of making a decision.

Measurement variables are critical to any empirical evaluation of CT. The use of CT skills is clearly the most difficult source of individual variability to measure. However, we may be able to employ methods used in the testing of dual-process theories to detect when CT is being used. For example, one indicator of System 2 processing is the inability to concurrently perform multiple tasks. Psychological and physiological indicators that are the indirect products of the application of CT skills may also be observed to determine when CT is engaged. Finally, concurrent verbal protocols taken while performing a task should reveal thoughts that are indicators of the application of CT skills. For example, an expression of thoughts or strategies about one’s own thinking reveals meta-cognition, which is an indicator of CT. Similarly, questioning the validity of the information given in the context or problem is an indicator of CT.

Process and outcome measures represent the intermediate and final indicators of the quality and effectiveness of CT, respectively. Process measures gauge how well the skills of CT, such as those listed in Appendix A, are executed. In contrast, outcome measures gauge how well the products of CT serve the relevant meta-task. Process measures directly assess an individual’s performance on CT skills. Since each skill would be executed to serve a meta-task, for example decision-making, the quality of its conduct as indicated by a process measure serves as an intermediate indicator of overall performance. Outcome measures, in contrast, are the final and cumulative indicators of performance.

The CT literature contains numerous examples of operationally-defined process measures related to CT skills. For example, the CT skill of “deductive reasoning” has been assessed through tests of syllogisms and logic tracing. In these cases, individual performance is typically measured by paper and pencil tests where participants receive a score based on the number of correct answers they provide. The way that particular skills are evaluated varies depending on the requirements of the skill. For example, the interpretative skill that describes the ability to “extract meaning” from a stimulus source has been assessed through multiple choice or essay tests of knowledge, as well as less direct markers that an individual has derived relevant material from a stimulus (e.g., behavioral or performance-based measures). Thus, a number of methods might be used to measure performance in executing a skill, and particular measures might be better for some meta-task contexts than for others. Similarly, outcome measures depend on the knowledge and skill domain being evaluated. A single measurement for all decision-making meta-tasks is not possible. Thus, the matter of defining particular process and outcome measures becomes one of operational definition in an empirical investigation.

VALIDATION OF THE CRITICAL THINKING MODEL

The model of CT described in the previous section changes the focus of CT research by specifying testable relationships among several variables. The CT model makes specific predictions about the effects of context variables such as ones that distinguish stimuli and tasks. It also specifies how predisposing individual difference factors and moderating subject factors affect CT. Affective, byproducts of CT are also specified. In summary, the CT model provides direction to future research that seeks to investigate initiating factors to, and products of, CT.

An experiment was conducted to examine some of the central predictions made by the CT model. The experiment described in this section was an investigation of the validity of the several of the CT model's key assertions. First, the CT model asserts that the situation must include substantive information, which is the material to which critical thinking skills are applied. This position is axiomatic to the CT model as one cannot think deeply about nothing. The CT model further asserts that substantive information will increase the tendency to employ CT skills compared to less substantive information. The CT model also states that the use of CT skills is more likely if the information presented by the situation is conflicting, disordered, uncertain, and complex or requires extensive logical reasoning. To test these predictions, the substance in stimulus material, as defined by the number of unique propositions, was varied. Additionally, one condition was examined in which inconsistent, i.e., contradictory, information was incorporated into the stimulus material.

A second set of predictions of the CT model tested by the experiment concerned the purposeful nature of CT. Specifically, the CT model states that one engages in CT skills only when the situational context includes at least one of four meta-tasks, i.e., when one's task is to understand some material, solve a problem, make a decision, or make an evaluative judgment. It was hypothesized that greater evidence of the application of CT skills should be observed when participants are asked to perform one of these tasks than when they are asked to perform a different task, such as sorting or identification. The current version of the CT model makes no assertions about the relative power of the four meta-tasks to elicit execution of CT skills. According to the CT model, therefore, judgment, understanding, decision-making, and problem solving should equally elicit CT. To test this hypothesis, participants were asked to perform three tasks (judgment, understanding, and identification of the general topic) on stimulus material that varied in substance.

A third prediction tested by the experiment involved the influence of predisposing individual difference factors on the application of CT skills. The CT model states that differences exist among individuals in their tendency to use CT. It was hypothesized that there is a positive relationship between independent measures of predisposition and indicators that CT has been used. However, the CT model makes no specific predictions about the relative influence of situational vs. predisposing factors. Thus, the relative strengths of predisposing and situational factors on CT are exploratory at this point. To evaluate the influence of predisposing factors, participants were asked to complete the Need for Cognition Scale (NFC) (Cacioppo, Petty, & Kao, 1984), an assessment instrument commonly regarded as an indicator of the tendency to engage in thought.

A fourth assertion of the CT model tested here involved the effects of moderating variables on CT. Specifically the CT model predicts that expertise or experience will affect the quality of CT, i.e., how well someone executes a CT skill. However, it should not affect whether one attempts to perform CT. To test this prediction, two groups of participants were recruited who varied in their degree of experience.

The fifth assertion tested by the experiment was that CT elicits negative affect because it requires greater effort. The CT model predicts that the application of CT skills should be associated with a corresponding increase in negative affect and increase in effort. Because the CT model predicts that various stimulus and task conditions should elicit the application of CT skills, it was hypothesized that these conditions will also elicit greater negative affect and require greater effort than conditions that are not hypothesized to elicit CT. This prediction was tested by measuring affect and effort immediately following each trial of the experiment.

In summary, the validation experiment tested five predictions from the CT model. The following method section provides specific information about the variables examined and measures employed in the experiment.

Method

Participants

Twenty-six participants (5 males and 21 females), ranging in age from 20 to 51 years of age, took part in the experiment. Three of the participants were recruited from Santa Barbara City College (SBCC), and 23 were recruited from the University of California, Santa Barbara (UCSB). All participants had taken a course in experimental psychology. Fifteen participants were enrolled in the graduate program in psychology at UCSB and eleven were undergraduate psychology majors. The undergraduates had completed an average of 3.1 years of college. The graduates had completed a mean of 7.1 years of college.

Participants were recruited through fliers posted in the psychology building at UCSB and through announcements made in class at SBCC. All participants were paid \$50.00 for their participation.

Materials

Each participant received a test booklet containing nine problems. Each problem presented the participants with a specific task to apply to a short paragraph that described a research investigation in the discipline of experimental psychology. Over the nine problems, participants were asked to perform three different tasks. The task instructions asked the participant to either 1) understand, 2) make a judgment about, or 3) simply identify the general topic of the material presented. The instructions for the three tasks are provided in Appendix C.

The substantive content of the paragraphs describing the nine research studies was also varied. Three different types of substantive content were presented. The first type contained very little substantive information, as measured by number of unique propositions presented. The second type was more substantive than the first, i.e., presented a greater number of propositions. The third type was as substantive as the second, but also included several

propositions that were inconsistent, i.e., contradictory, with one another. Thus, the third type offered substantive but degraded content. Each participant read three instances (descriptions of research studies) of the three types of substantive content, producing a total of nine problems. Nine research topics within experimental psychology were selected and low, high, and high/inconsistent paragraphs were developed for each topic. Thus, a total of 27 paragraphs were used, representing three content levels of each of nine research topics (see Appendix D). Table 2 shows the nine topics and the number of propositions for each content type.

Table 2. Number of Propositions per Topic and Substantive Content Type

Topic Paragraph	Substantive Content Type		
	Low Substance	Consistent High Substance	Inconsistent High Substance
Caffeine and Memory	12	21	21
Numerosity in Children	14	29	29
Daycare and Aggression	10	27	27
Girls' Sports Program Effectiveness	6	20	20
Hormone Replacement Therapy	11	22	22
Leading Questions and Memory	12	21	21
Nutrition Education	12	23	23
Groups Variables and Social Loafing	12	26	26
Valium and Phobia	12	23	23
Mean Number of Propositions	11.2	23.6	23.6
Standard Deviation	2.2	3.1	3.1

Participants also completed a response measure designed to assess the amount of mental effort expended to perform each problem. Effort was rated on a 5-point scale where 1 represented “Very Little Effort” and 5 represented “Extreme Effort”. Effort ratings were obtained after each of the nine problems by having the participants circle the appropriate level of effort on a separate piece of paper.

To measure the affect associated with the combined levels of task type and substantive content type, participants also rated how much they enjoyed completing each problem on three 7-point Likert scales. The three scales asked participants to indicate how pleasant the experience was (1 = Extremely Unpleasant and 7 = Extremely Pleasant), how enjoyable it was (1 = Extremely Unenjoyable and 7 = Extremely Enjoyable), and how positive or negative they felt while performing the task (1 = Extremely Negative and 7 = Extremely Positive).

The short form of the Need for Cognition Scale (NFC) (Cacioppo, Petty, & Kao, 1984) was also administered to each participant at the beginning of the experiment. The NFC scale asks respondents to indicate the extent to which 18 statements are characteristic of them. Participants used a 5 point scale (1 = extremely uncharacteristic and 5 = extremely characteristic) to evaluate each statement. For example, the NFC scale includes statements such as: “I find satisfaction in deliberating hard and for long hours” and “I really enjoy a task that involves

coming up with new solutions to problems”. The short-form NFC scale was used as a measure of predisposing attitude toward CT.

Design

The three levels of task type (understand, identify, and judge) and three levels of substantive content (low, high, high-inconsistent) were factorially combined to create nine conditions in a completely within-subjects design. The nine topic areas were crossed with the nine conditions of the experiment such that each participant read only one paragraph about each of the nine topic areas. A Latin Square was used to counterbalance the presentation order of the three levels of task type, general topic of the paragraph, and the three levels of substantive content. The dependent variables were time to complete each task, rating of effort for each task, rating of affect for each task, and indicators of CT quality from participants’ verbal protocols. In addition, the Need for Cognition Scale was used as a correlational measure.

Procedure

The experiment was conducted at the offices of Anacapa Sciences in Santa Barbara, California. Each participant was run individually, with one experimenter present at all times. Participants were first given an informed consent form that contained a written description of the experiment. Participants were then reminded that they were free to terminate the experiment at any time with no loss of compensation.

After informed consent was obtained, the experiment began by having participants complete the short form of the NFC scale. After completing the scale, participants were given a sheet of instructions describing the experimental procedures and their tasks. The experimenter went over the instructions verbally with each participant, after which, the participants completed a practice trial applying each of the three tasks to one practice paragraph. During the practice period, the experimenter answered questions and offered explanations of the three tasks. When the experimenter was satisfied that the participant understood the tasks, the experimental trials began. A video camera was used to record participants’ responses to the experimental trials.

Participants were given a stapled booklet that presented each problem task and paragraph on a separate page. Before beginning the first problem, participants were told to read through each paragraph completely and to talk through their reasoning as they worked out their responses to the task. They were also reminded that the experimenter would be unable to answer any questions about the paragraphs. When participants were finished with each problem, they turned the page to the mental effort and affect measures. When finished with these measures, the participants then turned to the next problem. The experimenter used a stopwatch to time the trials from the moment the participants began reading a paragraph to the moment they indicated they were through; separate times were recorded for each trial. After completing all nine problems, the participant was given a written debriefing statement explaining the experimental manipulations, and the experimenter answered questions about the investigation.

Measures. Dependent measures were recorded and analyzed, including the NFC scale, self-reported mental effort and affect ratings for each condition, response time for each trial, and indicators of the application of CT skills derived from the verbal protocols. Each of these

measures is relevant to one or more hypotheses generated by the CT model. In the context of this experiment, the NFC scale operationally defines a predisposition toward CT. The CT model presented in Section Two of this report predicts that NFC should be correlated with indicators of CT. The effort and affect ratings are operational definitions of the amount of mental work applied to each trial and the degree of enjoyment produced by each trial, respectively. As noted previously, the CT model predicts that reported effort should be greater for experimental conditions thought to elicit CT, i.e., those that require the tasks of judgment and understanding and those that require processing of high substance and/or inconsistent material. The CT model also predicts that these same conditions should produce lower affect ratings. Because the CT model states that CT takes more time than System 1 thinking, the conditions thought to elicit CT should also produce longer response times than other conditions of the experiment.

Two indicators of CT skills were derived from the verbal protocol data: (1) the number of questions of belief uttered by participants and (2) the number of times participants checked their thinking per condition of the experiment. Questions of belief were defined as statements that express skepticism about the validity of information presented in stimulus material. This measure operationally defines one feature of CT, which is carefully evaluating information to determine its truth value. The number of checks on one's own thinking is an operational definition of the meta-cognitive feature of CT. The CT model predicts that the experimental conditions thought to require CT (i.e., judgment and understanding tasks and high substance and/or inconsistent stimulus material) should produce a greater number of questions of belief and checks on thinking compared to the other conditions of the experiment. Further specification of each of the measures collected in this experiment is given below.

NFC. For the NFC scale, the scaling for several of the items was first reversed such that a high rating indicated positive need for cognition and a low rating indicated a negative need for cognition for all items. A single score for the NFC scale was then computed for each participant by summing the ratings (based on a five-point scale) of the 18 questions. The lowest possible NFC score was 18, which would indicate a low NFC, while the highest was 90, which would indicate a high NFC.

Effort. The raw self-reported ratings of mental effort (1=low effort, 5=high effort) provided by each participant for each of the nine conditions of the experiment were recorded and analyzed.

Affect. The relationships among the three affect scales (i.e., pleasant, enjoyable, negative/positive) were first analyzed to determine inter-item agreement. The raw ratings of the three affect items were strongly correlated, $r(233) = .84$ for Item #1 and #2; $r(233) = .80$ for Item #1 and #3; and $r(233) = .82$ for Item #2 and #3. Thus, a single affect score was computed by summing across each participant's responses to the three 7-point scales (minimum = 3, maximum = 21) for each of the nine conditions of the experiment.

Time Score. The absolute number of seconds taken by each participant to complete each of the nine trials of the experiment was recorded and analyzed.

Indicators of CT from Verbal Protocols. Each of the participants' verbal statements during the testing were transcribed from the video tape and then analyzed. The number of questions of belief participants asked was counted for each condition of the experiment.

Questions of belief were defined as utterances in which participants expressed doubt about whether he/she should believe information given in the summary paragraph being read. Thus, questions of belief were counted when participants criticized or doubted the conclusions or methods of an experiment. These included any comment that indicated the participant doubted the information given in the summary, wondered about missing information, or questioned the procedures of the experiment. Questions of belief also included any criticism or negative comment about the investigation's methods, results, analysis, etc. Questions such as "I wonder what it was like to run that study" were not counted as questions of belief.

Checks on one's own thinking operationally defined the meta-cognitive activity that is integral to the CT skills processed by System 2, as theorized by the CT model. The number of checks on thinking was counted for each condition of the experiment. Checks on thinking were defined as statements that referenced the participant's own thinking. For example, statements that express reflection on past or current thinking were counted, such as "I didn't really understand why they...but now I do". Statements about strategies a participant planned to take were also counted such as "I'm kind of reasoning through this" and "I'm going to read it again." Finally, going back and rereading portions of the statements were also counted in this measure because they were considered evidence of recognition that the level of understanding obtained on the first read-through was not adequate.

Analysis of the verbal protocol data was performed first by a single rater. A second rater then analyzed 20% of the protocols to evaluate the inter-rater reliability of the ratings. Inter-rater agreement for both measures (i.e., number of questions of belief and number of checks on thinking) was high ($r = .974, p < .05$; $r = .969, p < .05$, respectively), indicating that these measures are highly reliable.

Each of the dependent measures of participant response to the nine conditions was analyzed separately. The results of these analyses, presented below, are organized by the hypothesis tested and dependent variable measured.

Results

Effect of Task and Stimulus Variables on Indicators of CT

As noted previously, several measures taken in the experiment were operational measures of the use of CT skills including trial completion time, self-reported effort, the number of questions of belief, and the number of checks on thinking. If the CT model is correct in stating that System 2 is a relatively slow process, then conditions hypothesized to elicit CT should take longer than other conditions. Similarly, greater effort should be reported for those conditions hypothesized to produce CT. With regard to measures taken from the verbal protocols, more questions of belief should be observed in conditions hypothesized to elicit CT according to the CT model. For each of these measures, the CT model predicts that the understanding and judgment tasks should elicit CT more often than the identification task. Similarly, the CT model predicts that the high substance and high substance/inconsistent stimulus conditions should elicit CT more often than the low substance condition. The results of the analyses of each measure are reported below.

Time. The number of seconds taken by each subject to complete each trial of the experiment was submitted to two-way repeated measures ANOVA comparing the effects of variation in stimulus material and task. Three participants were excluded from the analysis due to missing time data. The results indicated an overall effect of substantive material, $F(2, 44) = 18.28, p < .001$. Post hoc analyses using a Bonferroni adjustment for multiple comparisons indicated that material that is highly substantive but contains inconsistent material took longer to process ($M = 511, SD = 249$) than substantive consistent material ($M = 390, SD = 115$), $p = .013$. The latter material took longer to process than did low substance material ($M = 317, SD = 104$), $p < .001$.

Significant differences in response time were also produced by the task variable $F(2, 44) = 11.97, p = .001$. Post hoc analyses with Bonferroni adjustments showed that the understanding and judgment tasks took comparable amounts of time in seconds to complete ($M = 449, SD = 190$ and $M = 471, SD = 191$, respectively). However, they both took longer than identifying the general topic of the paragraphs ($M = 291, SD = 109$), $p < .001$ for both.

Time, in seconds, taken to complete the trials was also influenced by the combined effect of task and substance of the stimulus material, ($F(4, 88) = 2.66, p < .04$). Figure 2 shows that the degree of substance of the material has little effect on the time it takes to discern the general topic of a summary of an experimental investigation. However, high substance material that is inconsistent takes longer to understand and judge than high substance consistent material, which in turn takes longer than low substance material.

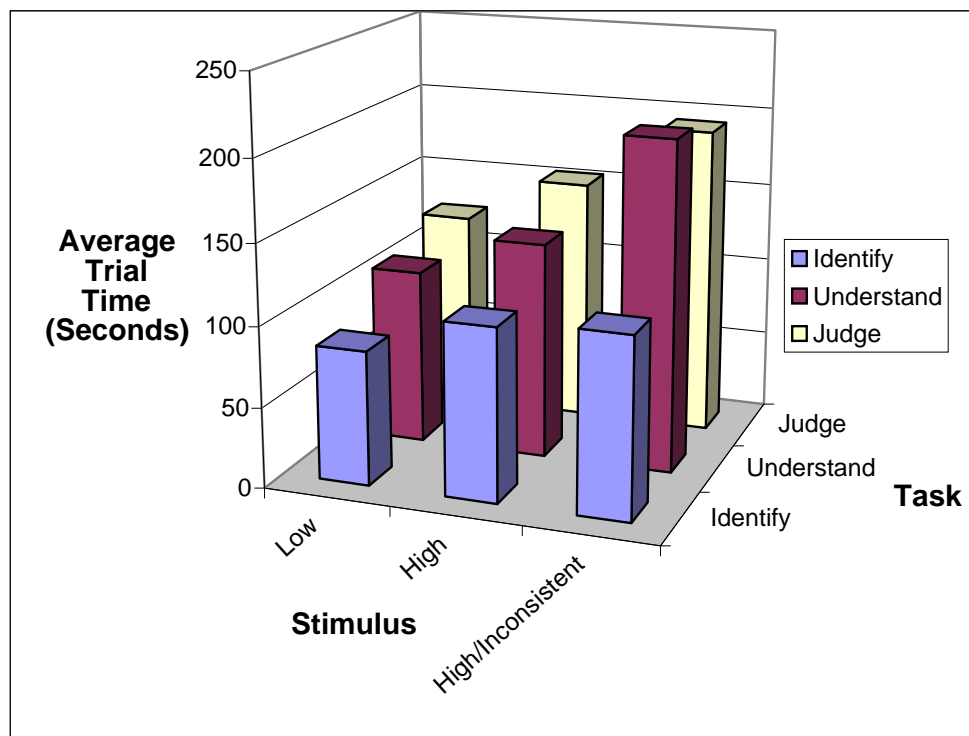


Figure 2. Mean response time as a function of substance of stimulus material.

Effort. The self-reported ratings of effort for each condition of the experiment were submitted to two-way repeated measures ANOVA comparing the effects of variation in stimulus material and task. One case was deleted from the analysis due to missing data. The effects of stimulus material and task observed in the response time data were again seen in the ratings of effort. Participants rated the high substance material ($M = 3.05$, $SD = .60$) and the high substance/inconsistent material ($M = 3.25$, $SD = .65$) as more effortful to process than the low substance material ($M = 2.37$, $SD = .75$), $F(2, 48) = 21.633$, $p < .001$. Post hoc analyses revealed that the two high substance conditions were rated as requiring comparable level of effort.

Participants also rated the three tasks as differentially effortful, $F(2, 48) = 11.36$, $p < .001$. The understanding and judgment tasks were rated as equally effortful ($M = 3.01$, $SD = .58$ and $M = 3.19$, $SD = .61$, respectively). Post hoc comparisons showed they were both rated as more effortful than identifying the general topic of the paragraph ($M = 2.48$, $SD = .86$, $p < .001$).

As in the response time data, a statistically significant interaction was observed in the ratings of effort, $F(4, 96) = 2.58$, $p < .05$. When the task is to identify the general topic of the paragraph, the substance of the material has little to no effect on how effortful participants perceive the task. However, if the task is to make a judgment or to understand the material, more substantive material and inconsistent material are considered more effortful to process. Figure 3 shows the interaction between task and substance of the material on ratings of effort.

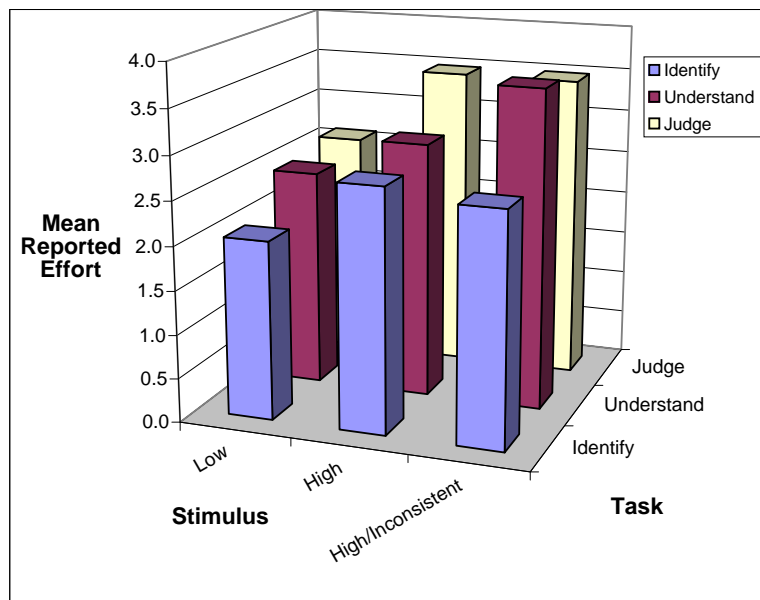


Figure 3. Mean reported effort as a function of task and substance of stimulus material.

Number of Questions of Belief. The number of questions of belief uttered by participants for each condition was submitted to a factorial ANOVA comparing the effects of task type and substance of stimulus material. The potential moderating effects of a third variable, level of

education, were also investigated in this analysis. However, the results of education level are reported later in a separate subsection.

The main effect of the substance of the material read was statistically significant, $F(2, 44) = 10.43, p < .001$. Participants asked more belief questions about high substance but inconsistent material ($M = 1.54, SD = .89$) than either of the other two conditions. However, post hoc comparisons using a Bonferroni correction indicated that a comparable number of questions was asked about low and high substance material ($M = .92, SD = .77$ for low substance; $M = .65, SD = .64$ for high substance).

The analysis of the task variable indicated that subjects asked more questions when they had to make a judgment than if they had to simply understand the material. The latter condition, in turn, generated more questions than when participants identified the general topic of the material $F(2, 44) = 25.65, p < .001$. Post hoc analyses using a Bonferroni correction for multiple comparisons revealed significant differences between the judgment and understanding tasks, and between the understanding and topic identification task. On the average, participants asked 1.9 ($SD = 1.0$) questions per problem on judgment tasks, but asked only .88 questions ($SD = .53$) when they had to understand the material. They asked only .40 questions ($SD = .62$) when they had to discern the general topic of the material.

The significant interaction between substance of material and task ($F(4, 88) = 6.90, p < .001$) revealed a more complex picture of the factors that affect questioning of beliefs. Figure 4 depicts the mean number of belief questions asked in each of the nine tasks by substance conditions. When participants were asked to identify the general topic of the material they read, the substance of the material had no effect on the number of belief questions that they asked which was very few. When the task was to understand the material, however, a greater number of questions were asked if the material was highly substantive, but inconsistent. When participants were asked to make a judgment, they asked more questions about paragraphs with inconsistent information and paragraphs with little substance than of paragraphs that had consistent high substance. There was no difference between low substance and inconsistent paragraphs in the judgment condition.

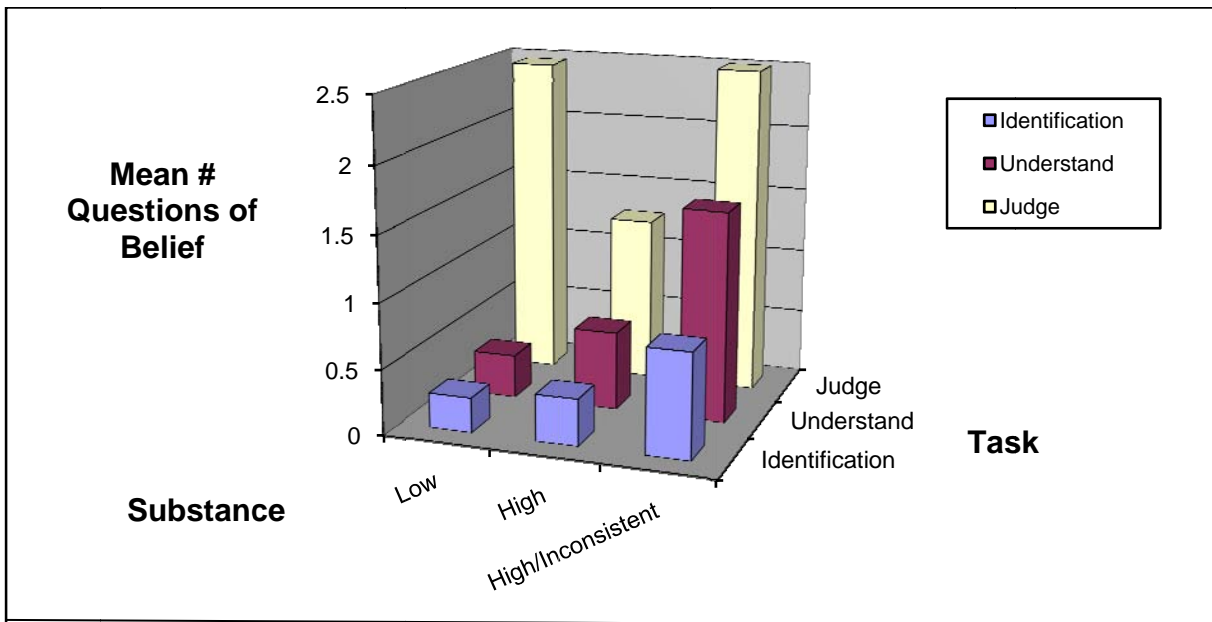


Figure 4. Mean number of questions of belief as a function of task and substance of stimulus material.

Number of Checks on Thinking. The number of checks on thinking derived from the protocol data for each condition of the experiment was submitted to a three-way factorial ANOVA comparing the effects of task type, stimulus substance, and the potential moderating effects of a third variable, level of education. The effects of educational level are discussed later in a separate subsection.

The results of the analysis indicated that participants checked their thinking more frequently when asked to make a judgment about a summary of an experiment than when asked to identify the general topic, $F(2,46) = 4.18, p < .02$. Post hoc analyses showed that the mean number of checks on thinking per trial produced by the understanding task ($M = .50, SD = .78$) was not significantly different from either the number produced by the identification task ($M = .31, SD = .40, p > .05$) or from the judgment task ($M = .56, SD = .72, p > .05$).

The degree of substance of the material also affected the number of checks on thinking, $F(2, 46) = 7.19, p = .002$. Post hoc analyses revealed that fewer checks per trial were performed when subjects read low substance summaries ($M = .21, SD = .66$) than when they read inconsistent substance summaries ($M = .79, SD = .75$), $p < .01$. However, there was no reliable difference between the number of checks produced by low substance and high substance summaries, $p > .05$. Inconsistent high substance summaries also produced a significantly greater number of checks on thinking than did consistent high substance summaries ($M = .36, SD = .53$), $p < .01$. The task and substance variables had additive effects on the number of checks on thinking; no interaction was observed.

Effect of CT on Affect

To determine if the conditions of task and stimulus substance hypothesized to elicit CT were also associated with increased negative affect, the summed ratings of affect were submitted to a two-way (stimulus substance and task) repeated measure ANOVAs. Participants differentially enjoyed summaries that varied in substance, $F(2, 50) = 3.21, p < .05$. Post hoc analyses showed that participants reported lower affect for summaries that included inconsistent propositions ($M = 13.6, SD = 2.3$) compared to either high or low substance summaries ($M = 14.5, SD = 2.2$ for high substance; $M = 14.6, SD = 2.0$ for low substance). The omnibus ANOVA indicated that participants also rated the three tasks differentially enjoyable, $F(2, 50) = 3.24, p < .05$. However, post hoc comparisons revealed only marginal statistical significance. The judgment task was rated as more enjoyable than either the understanding task or the identification task, but only at marginal levels of significance ($p < .05$) comparing identification to judgment; $p < .07$ comparing understanding to judgment (all uncorrected for multiple comparisons). No difference in affect was observed between the identification and understanding tasks. The hypothesized interaction between stimulus substance and task was not observed in the affect data.

Effect of Moderating Variable (Level of Education) on CT

The effect of level of education on several measures was analyzed. First, to determine if the undergraduate and graduate student groups were different in ways other than educational level, their NFC scores, overall response time, reported affect, and reported effort were compared. To determine if educational level played a moderating role in eliciting CT, it was analyzed in the context of a factorial ANOVA that also analyzed the combinatorial effects of stimulus substance and task on the number of questions of belief expressed by participants. We discuss the findings of each of these analyses in this section.

Level of Education and NFC. Level of education was not related to NFC. Graduate students and undergraduates exhibited about the same level of NFC, $p > .05$.

Level of Education and Response Time, Affect, and Reported Effort. No significant differences were observed in the time undergraduates and graduates took to complete the trials of the experiment, $p > .05$. Also, no significant differences were observed in the affect ratings reported by undergraduates and graduates, $p > .05$. Level of education did not affect reported effort either. Graduate students and undergraduates reported similar levels of effort for all conditions of task and stimulus substance, $p > .05$.

Level of Education and Questions of Belief. Level of education produced a marginally significant effect on questions of belief ($F(1, 22) = 4.24, p < .05$) indicating that graduate students asked more questions of belief over all conditions ($M = 1.2, SD = .3$) than did undergraduates ($M = .8, SD = .5$). Level of education also interacted with the effects of substance and task to determine the number of questions of belief posed by participants. Figure 5 shows that undergraduates were less affected by variation in substance than were graduate students, $F(2, 44) = 3.75, p < .03$. Graduate students, in contrast, asked more questions when processing inconsistent material. Similarly, graduate students differentially reacted to variation in task, whereas undergraduates did not, $F(2, 44) = 7.75, p < .001$. Figure 6 shows that graduate

students asked more questions of belief when they were tasked with making a judgment than with understanding or identifying. In contrast, undergraduates asked roughly the same number of questions for all three task conditions.

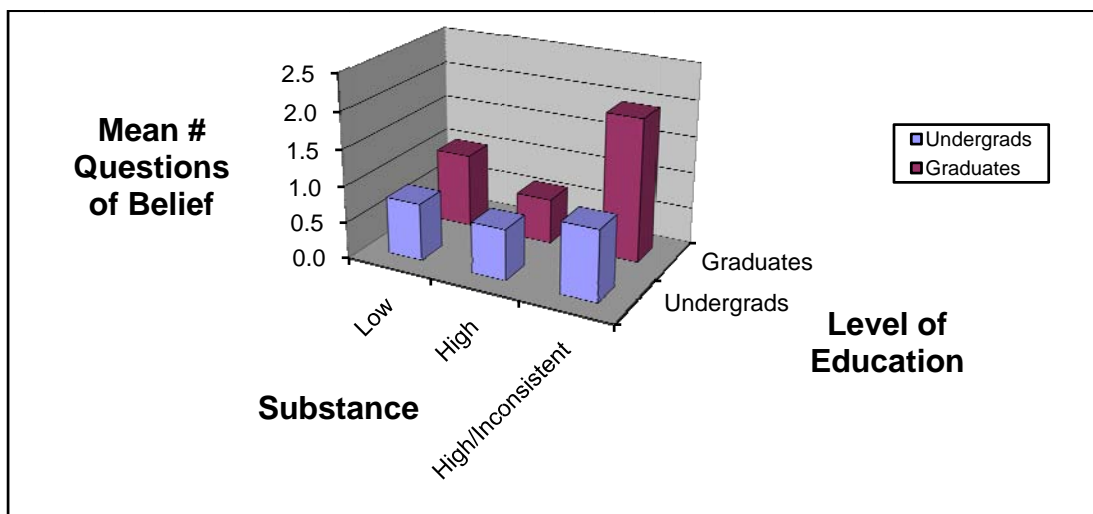


Figure 5. Mean number of questions of belief as a function of substance of stimulus material and educational level.

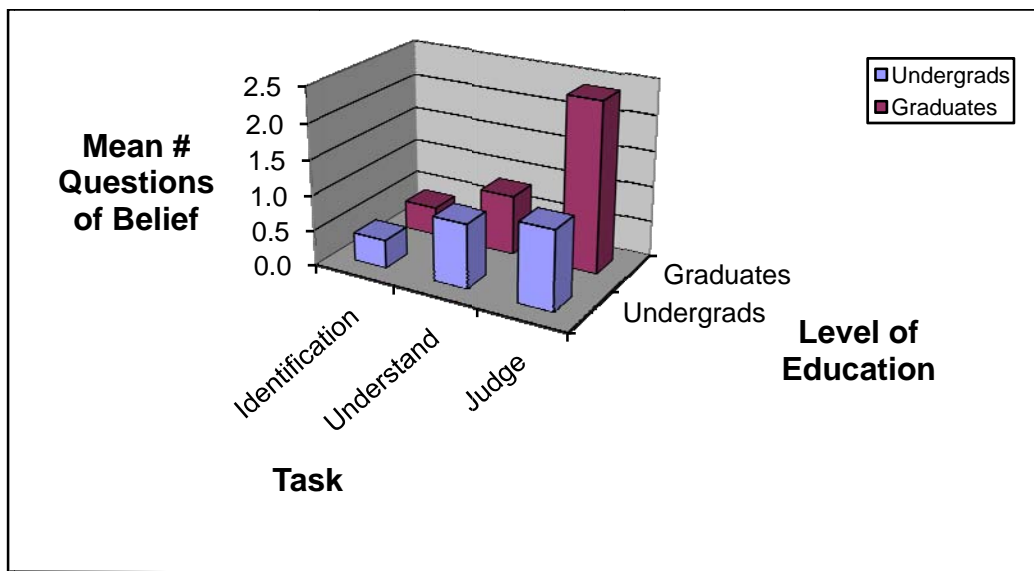


Figure 6. Mean number of questions of belief as a function of task and educational level.

Figures 7a and 7b show the marginally significant three-way interaction between stimulus substance, task, and level of education, $F(4,88) = 3.24$, $p < .07$. Taken together, these figures suggest that graduate students ask more questions when making judgments about inconsistent

and low substance material than in any other condition of the experiment, whereas undergraduates do not ask many questions about inconsistent material.

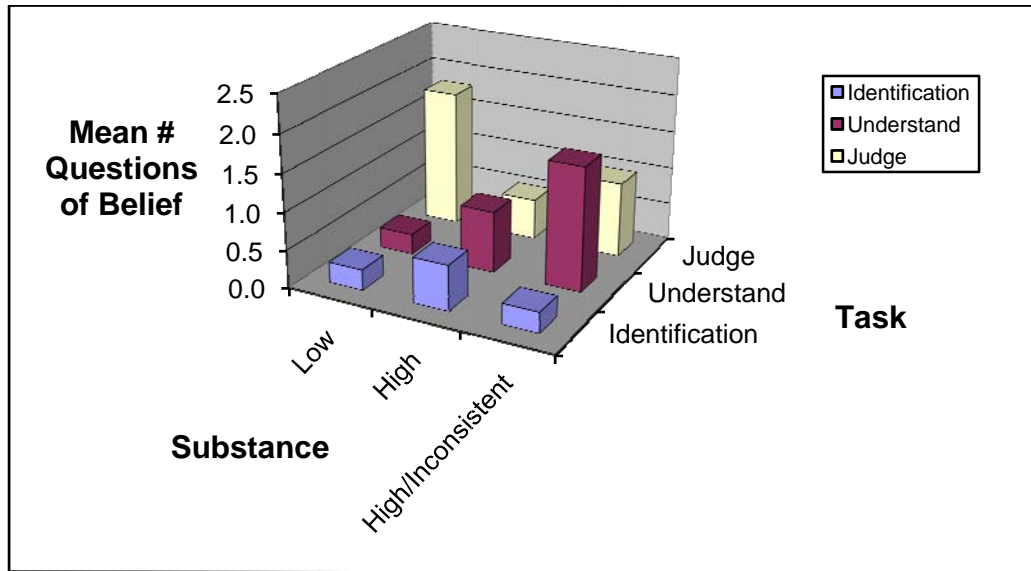


Figure 7a. Mean number of questions of belief asked by undergraduates as a function of task and substance of stimulus material.

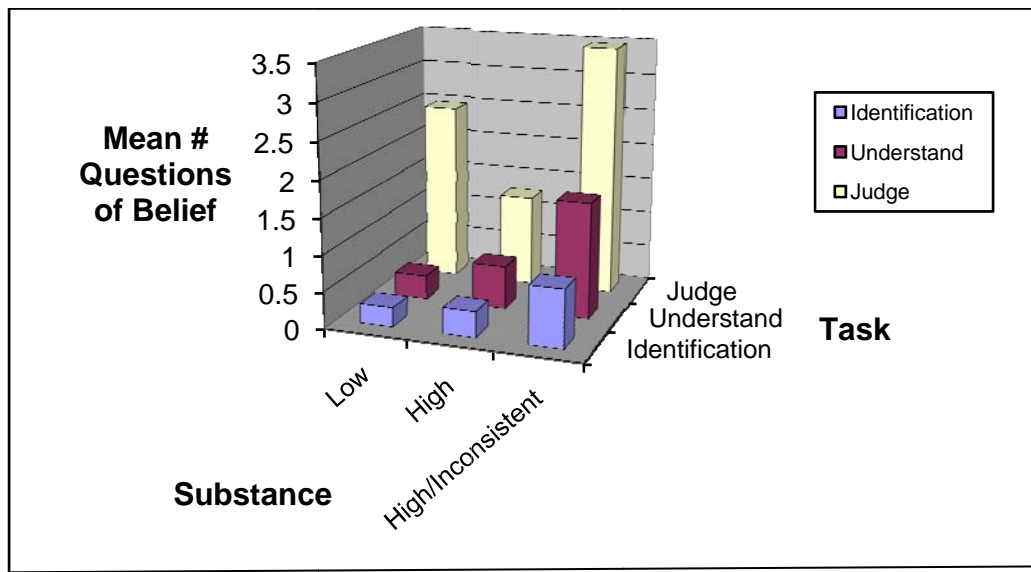


Figure 7b. Mean number of questions of belief asked by graduate students as a function of task and substance of stimulus material.

Level of Education and Checks on Thinking. While there was no overall statistical difference between graduate and undergraduate students in the number of checks on thinking they performed, the two groups of participants responded differently to the task variable. This was shown by a significant interaction between education level and task, $F(2, 46) = 4.17, p < .02$. Figure 8 shows that graduate students performed the most checks on their thinking when asked to make a judgment. In contrast, undergraduates checked their thinking most when asked to understand the material, and checked very infrequently when asked to either identify the general topic of the summary or to make a judgment.

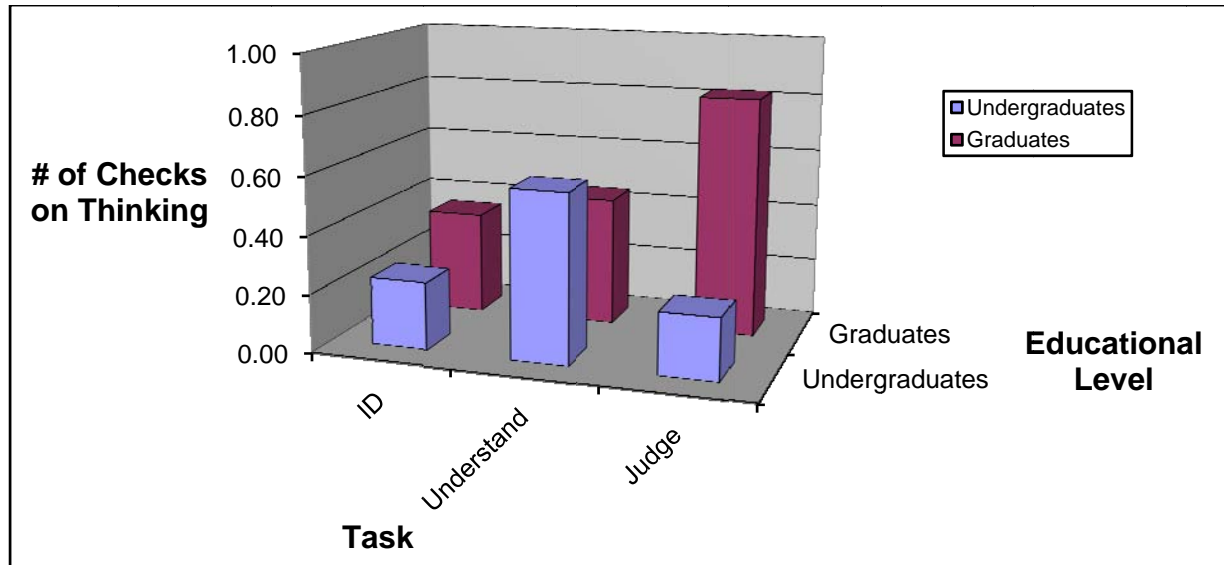


Figure 8. Mean number of checks on thinking as a function of task and educational level.

Relationship of Predisposing Factor (Need for Cognition) to CT

The NFC scale scores were correlated with a variety of other measures to determine if NFC plays a moderating role in controlling CT. Table 3 provides descriptive statistics on the NFC scale and each of the measures. Also shown in Table 3 are Pearson Product Moment correlations that describe the relationship between NFC and several measures of response time, reported effort, reported affect, and the number of questions of belief. Bonferroni probabilities were also computed for each correlation. NFC did not significantly predict any measure of time, reported effort, reported affect or questions of belief, with $p > .10$ in all cases. The reported r values in Table 3 all fail to reach statistical significance. In summary, no evidence was obtained showing that this particular potential predisposing factor (NFC) influenced performance in this experiment.

Table 3. Sample Descriptive Statistics for Need for Cognition and Other Measures

Measure	Correlation with NFC (<i>r</i>)	Mean	Range	Standard Deviation
NFC (score)		72.4	37	7.9
Total Response Time for All 9 Conditions (sec)	-.14	1215.5	1505	390.7
Total Time for All Identification Trials (sec)	-.31	291.9	379	109.0
Total Time for All Understanding Trials (sec)	-.03	449.6	826	190.0
Total Time for All Judgment Trials (sec)	-.08	471.5	832	191.1
Total Time for All Low Substance Stimulus Trials (sec)	-.31	317.4	475	104.4
Total Time for All High/Consistent Substance Stimulus Trials (sec)	-.22	390.0	524	115.6
Total Time for All High/Inconsistent Substance Stimulus Trials (sec)	-.03	511.0	831	249.4
Average Affect for All Trials	.15	14.3	6.8	1.8
Average Affect Identification Trials	.08	13.9	8.3	2.2
Average Affect Understand Trials	.11	14.0	8.3	2.2
Average Affect Judgment Trials	.18	14.9	7.0	2.1
Average Affect Low Substance	.09	14.6	7.0	2.0
Average Affect High/Consistent Substance	.30	14.5	9	2.2
Average Affect High/Inconsistent Substance	-.01	13.6	8	2.3
Average Effort for All Trials	-.38	2.8	2	.5
Average Effort Identification Trials	-.36	2.4	3.0	.8
Average Effort Understand Trials	-.22	3.0	2.3	.5
Average Effort Judgment Trials	-.29	3.1	2.6	.6
Average Effort Low Substance	-.22	2.3	2.6	.7
Average Effort High/Consistent Substance	-.35	3.0	2.3	.5
Average Effort High/Inconsistent Substance	-.36	3.2	2.3	.6
Average # Questions for all Trials	.03	1.0	2.2	.5
Average # Questions Identification Trials	.04	.4	2.3	.6
Average # Questions Understand Trials	-.10	.8	2.0	.5
Average # Questions Judgment Trials	.08	1.8	3.3	1.0
Average # Questions Low Substance	.10	.9	3.3	.7
Average # Questions High/Consistent Substance	-.23	.7	3.5	.8
Average # Questions High/Inconsistent Substance	.19	1.5	3.3	.8

In interpreting the lack of significance above, it appears that the NFC scores for the sample of individuals who participated in this experiment may be restricted in range. The minimum score possible for the short form of the NFC is 18 and the maximum is 90. However, the lowest NFC score produced by the sample was 49 and the highest was 86, showing that most scores fell in the upper half of the potential range. An individual who consistently responded to the NFC items by answering “uncertain” or “neither characteristic nor uncharacteristic” would receive a score of 54. Therefore, nearly all participants scored in the upper half of the potential range of scores for the NFC, indicating a positive need for cognition. Moreover, 68 percent of the sample produced NFC scores between 68 and 80. Given that all subjects in the experiment were college students or graduate students, such high average NFC scores is not surprising. The experiment would need to be replicated with a broader sample of subjects (e.g., some with high school only or less) to see an effect of NFC.

Discussion

Several of the hypotheses generated by the model of CT presented in the previous section were supported by the present experiment. However, the results of the experiment also suggest that further refinements should be made to the CT model. The discussion of the findings of the validation experiment is organized by the hypotheses generated by the CT model.

Effect of Substance of Material on CT

As previously discussed, the CT model asserts that substantive information will increase the tendency to employ CT skills compared to less substantive information. A second prediction is that the use of CT skills is more likely if the information presented is conflicting, disordered, uncertain, complex or requires extensive logical reasoning. It is predicted that high substance material and inconsistent material should take longer to process, require more effort, generate more questions of belief, and generate more checks on thinking than stimulus material that is relatively low in substance.

These predictions were generally supported by the response time and reported effort measures. Substantive material took longer and more effort to process than low substance material. Response times were longer for inconsistent substantive material than consistent substantive material; however, they were not rated as more effortful. Moreover, these effects occurred only when the task was to understand or judge the summaries, which is predicted by the CT model. When the task is not one of the predicted purposeful meta-tasks, substance of the material has little effect on time or effort. When considered alone, the effects of variation in stimulus substance on these two measures supports the CT model's predictions.

However, a more complex picture emerges when one examines the effects of stimulus substance on questions of belief and checks on thinking, which are arguably better indicators of CT. Only inconsistent, high substance material generated more questions of belief than low substance material. High substance consistent material generated about the same number of questions of belief as low substance material. Similarly, more checks on thinking were produced by inconsistent material than by low substance material. High consistent material generated about the same number of checks on thinking as did low substance material.

Examining all four measures, these results do not support the prediction that high substance material increases the tendency to apply CT skills. The interaction between task and substance on questions of belief further fails to support the CT model's prediction concerning substance. It shows that under some task conditions, low substance material can actually generate *more* CT than high substance material. Specifically, when asked to make a judgment about low substance material, more questions of belief were asked than when judging high substance material.

The results, however, do support the prediction that the application of CT skills is more likely if the available information is degraded in some way, i.e., is conflicting, disordered, uncertain, etc. A general pattern emerges that inconsistent material tends to take longer, and produces more questions of belief and checks on thinking.

It is possible that the substance of material was not effectively manipulated in this experiment. That is, the number of propositions in a stimulus may not be an adequate operational definition of substance or there may not have been enough of a difference between the number of propositions in the low and high substance summaries to realize an effect. Thus, future research should address the manipulation of substance on the tendency to apply CT skills.

These results have a practical implication for the design of information systems and for training that seeks to increase the tendency to apply CT skills. Designers should be aware that it is possible that people may not question or check highly substantive consistent material any more than low substantive material. If CT is desired, inconsistent content might be highlighted by information systems. Similarly, if training systems seek to encourage CT, one strategy would be to sensitize students to inconsistent material.

In summary, the results suggest that the CT model's assertion that increasing substance of the material is positively correlated with the application of CT skills should be revised. Or the experiment might be replicated using a revised operational definition of substance. However, the results are consistent with the proposition that inconsistent material increases the likelihood that CT skills will be engaged.

Effect of Task on CT

According to the CT model, CT skills should only be engaged when the situational context includes at least one of four meta-tasks. As noted previously, the CT model makes no assertions about the relative power of the four meta-tasks to elicit execution of CT skills. Therefore, judgment and understanding tasks should equally elicit CT.

The response time and effort ratings supported the predictions concerning task. Both the judgment and understanding tasks took longer and were more effortful than the identification task. The prediction was also supported by the questions of belief measure in that both the judgment and understanding tasks generated more questions of belief than did the identification task. Similarly, the judgment task produced more checks on thinking than did the identification task. However, the understanding task failed to generate more checks on thinking than the identification task. Therefore, the prediction that understanding is a meta-task that elicits CT was not supported by the checks on thinking data. Moreover, it appears that understanding and judgment may not equally elicit CT. The finding that the judgment task produced more questions of belief than the understanding tasks suggests that the four meta-tasks are not necessarily equal in their power to elicit CT.

These results suggest that refinement of the CT model may be needed with regard to task. While they generally support the idea that understanding and judgment encourage the application of CT skills, they also suggest that judgment may be more effective at doing so. Perhaps participants are more motivated to critically think when the consequences of their actions are higher, as would be true when making a judgment as compared to simply understanding. While the performance produced in both tasks could be questioned, the judgment task may have more important consequences. This was certainly true in the operational definition of task used in this experiment where participants had to judge whether or not to accept a manuscript for

publication, as opposed to the understanding task where participants had to prepare for a test. While both manipulations had implied consequences, they were more severe for the judgment task. This argument suggests that the consequences of tasks may play a part in encouraging CT. Future empirical research and refinement to the CT model should focus on the effects of task consequences.

Effect of Predisposing Factors on CT

The CT model states that differences exist among individuals in their tendency to use CT. One should observe a positive relationship between independent measures of predisposition and indicators that CT has occurred. The results of this experiment failed to support the notion that predisposing individual difference factors affect the tendency to engage in CT skills. The NFC scale failed to correlate with any measure of response time, effort, affect, questions of belief, or checks on thinking. However, potential restriction of range problems on the NFC scale may have produced this null effect. At this point, it is premature to conclude that predisposing factors do not affect the application of CT skills. However, within the context of this particular experiment, the experimental conditions largely determined response time, effort, affect, questions of belief, and checks on thinking. It is possible that the effects of task and stimulus substance were sufficiently strong so as to overwhelm any effect that predisposing factors may have had. One might even question whether they produced a sort of demand characteristic, influencing subject protocols and other measures of behavior. Or the NFC Scale may not be a strong measure of predisposition for critical thinking. To resolve these issues, future research should focus on examining other measures of predisposing factors under other environmental conditions.

Effect of Moderating Variables on CT

According to the CT model, expertise and experience should affect the quality of one's CT, but not the likelihood that one will engage in CT skills. The present experiment tested the prediction that experience level, which was operationally defined as education, should affect how the CT skills are applied. Specifically, graduate students should take more time to complete trials, report more effort, and generate more questions of belief and checks on thinking as undergraduates.

The CT model's prediction concerning experience was not supported in the response time and effort data. No differences were observed between graduates and undergraduates in these two measures, nor was there any difference in reported affect between the two groups of participants.

The predicted effect of experience was supported by the more decisive indicators of CT: questions of belief and checks on thinking. Graduates asked more questions of belief, although this was only a marginally significant difference. The number of questions of belief generated by graduates was also more sensitive to stimulus and task variables than the number asked by undergraduates. The number of questions asked by graduates was particularly affected by the combined condition of judgment task and inconsistent material, and the combined condition of judgment task and low substance material. It appears that graduates are skeptical about material if they have to make a judgment when that material has little or inconsistent content. Graduates

also checked their thinking more when asked to make a judgment. Undergraduates asked fewer questions in general. However, they tended to ask relatively more questions when they had to understand high substance material that was consistent, and when they had to judge low substance material, than in the other conditions.

These findings are consistent with the idea that experience does affect the application of CT skills. Several components of experience may be the source of the effect. It may be that experience increases knowledge, which then tends to increase both the ability to process information and the ability to recognize weaknesses in the information. Alternatively, it may be that experience teaches caution and skepticism. Yet another potential explanation is that graduates and undergraduates are different from one another in more ways than just experience, which could account for their differences in questioning and checking. Selection bias may be working to create these differences. For example, graduate students may be more intelligent or more curious, each of which might increase the number of questions they ask and the checks they perform on their thinking. However, the graduates who participated in the present experiment were not different from the undergraduates in NFC, reported affect, reported effort, nor response time, which one would not expect if they were differentially intelligent or curious.

In summary, the findings of the present experiment suggest that educational level plays a part in affecting the application of CT skills. There are, however, many covariates of educational level, and the one responsible for the observed differences between graduate and undergraduate students cannot be determined by this experiment. Nonetheless, the findings indicate that increased experience is positively related to the increased application of CT skills. Future research should focus on isolating the individual difference variable that makes graduate students more likely to apply CT skills.

Effect of CT on Negative Affect

According to the CT model, the application of CT skills should be associated with a corresponding increase in negative affect. Those conditions hypothesized to elicit CT, such as the tasks of understanding and judgment and high substance and inconsistent material, should be rated as less enjoyable than the other conditions of the experiment.

The findings of the present experiment were mixed in their support of this hypothesis. Participants reported that they enjoyed processing inconsistent summaries less than the other summaries, which is consistent with the CT model's prediction. However, the equivalent ratings for the low and high substance material failed to support the prediction. The task that should be associated with lower affect according to the CT model, judgment, was actually rated as more enjoyable to perform than understanding or identification. Although other results of the experiment indicated that judgment did elicit CT, it appears that participants found that experience enjoyable.

In summary, negative affect was not a simple, direct outcome of CT in this experiment. Instead, it appears that other factors determined the level of enjoyment experienced by participants. The findings suggest that CT does not always elicit negative affect, and that the CT model should be revised accordingly. Additional research is necessary to uncover the

determinants of negative affect and its relationship to CT, and to determine the conditions under which negative affect is, and is not, produced.

Summary and Conclusions

The model of CT described in this report generated a number of predictions that previously had not been empirically tested. The CT model was sufficiently specified to permit falsification of many of its assertions. The present experiment tested five of the model's central predictions. As a result, we now have a clearer picture of the effects of CT on judgment and understanding. A better understanding of the effects of stimulus substance on CT. In addition, the results show that clear CT does not always generate negative affect and that experience may well increase the likelihood and quality of CT.

Although the results of the validation experiment were mixed in their support of the CT model, the CT model has passed an important scientific criterion. It has generated testable hypotheses. This experiment provides one test of a portion of the CT model's predictions; there are many more hypotheses to be tested and the findings of the present experiment should be examined, replicated, and extended. Some of the findings point to places in the CT model that require greater specification or modification. Other findings are consistent with the CT model's predictions. Future research is needed to further develop the CT model and to increase our understanding of CT.

AN INVESTIGATION OF CRITICAL THINKING IN ARMY BATTLE COMMAND

The results of the validation experiment supported some of the CT model's predictions. However, confirmation of predicted relationships generated by a model is only one component of its validation. The degree to which any psychological model can be applied to real-world situations is also an important measure of its validity. The ecological validity of our model of CT is particularly important to its ability to guide the design of training that would improve CT skills. If the CT model fails to capture important variables relevant to CT in performing practical tasks, its ability to focus training efforts will be handicapped. Thus, a second investigation was conducted to determine whether the model of CT could be applied to the domain of Army battle command. The domain of Army battle command was chosen because it demands high levels of CT ability for reasons previously discussed.

The central focus of the investigation was to determine whether the CT model adequately defined CT skills, situational conditions, and predisposing factors important and/or problematic to Army battle command. A survey instrument was developed and administered to a sample of Army officers to address these questions. After-action interviews that focused on survey responses were also conducted to obtain a better understanding of survey responses. The methods and results of the investigation are described below.

Method

Participants

Eighteen Army officers stationed at Fort Hood, Texas participated in the investigation during April, 2000. Three lieutenant colonels, eight majors and seven captains participated in the investigation. The officers primarily served in the infantry and armor branches of the Army, and most were currently engaged as staff officers at the battalion or brigade echelons. Table 4 shows the ranks, duty positions, and branches of the 18 participants.

Survey Materials

Participants completed a five-page survey that assessed their opinions and experiences concerning CT skills, predisposing attitudes, and situational conditions as applied to the domain of battle command. The section of the survey that addressed CT skills asked the officers to evaluate thirteen broad classes of skills⁷ organized into three major types (interpretive, reasoning, and meta-cognitive skills). The officers were asked to rate the importance of each of the thirteen classes of skills for battle command on a 4-point scale, with anchors: very important, important, somewhat important, and not important. They were also asked to indicate if, in their experience, they had observed problems in the execution of each class of skill. In addition, survey participants were asked to identify one or more battle command tasks that required execution of each skill. The thirteen broad categories of CT skills are shown in Table 5.

⁷ It was necessary to reduce the large number of CT skills to a manageable number that could be reasonably evaluated by participants within the time constraints of a survey and interview. Thus, the 130 CT skills shown in Appendix B were broadly classified into 13 categories.

Table 4. Rank, Duty Position, and Branch of Participating Officers

Rank	Current Duty Position	Branch	# Officers
Lieutenant Colonel	Battalion Commanding Officer	Infantry	1
Lieutenant Colonel	Battalion Commanding Officer	Armor	1
Lieutenant Colonel	Brigade Executive Officer	Infantry	1
Major	Brigade Operations	Armor	1
Major	Battalion Executive Officer	Infantry	1
Major	Battalion Operations Officer	Infantry	2
Major	Battalion Operations Officer	Armor	4
Captain	Company Commander	Infantry	3
Captain	Company Commander	Armor	1
Captain	Brigade Assistant Operations	Infantry	1
Captain	Brigade Logistics Officer	Infantry	1
Captain	Battalion Intelligence Officer	MI	1
Total			18

The importance of predisposing attitudes to battle command was similarly assessed by having the officers rate each of nine individual difference factors on the importance scale described above. Eleven situational conditions were similarly assessed. The survey was prefaced with instructions and a description of CT written in non-technical language. Officers were also asked to provide information about their work experience and background.

Procedure

The topic of CT was introduced and the purpose of the investigation was described to the participants. Officers were then given a copy of the survey of thirteen broad categories of battle command tasks for each CT skill. Officers were instructed to identify problematic CT skills as those that were (1) not performed, (2) not performed well, (3) used indiscriminately, or (4) performed excessively. After completing the CT skill survey items, participants were asked to complete the predisposing attitudes and situation condition sections of the survey.

Table 5. Means and Distribution of Importance Ratings and Frequency of Problems for 13 Broad Classes of CT Skills

Critical Thinking Skill	Problematic	Very Imp.(1)	Imp.(2)	Somewhat Imp (3)	Not Imp. (4)	No. Resp.	Mean
INTERPRETATION SKILLS							
Framing the problem	10	14	4	-	-	18	1.22
Extracting meaning from the material	8	10	5	3	-	18	1.61
Assessing materials for consistency, clarity, and completeness	7	9	8	1	-	18	1.56
REASONING SKILLS							
Inductive reasoning	9	14	3	1	-	18	1.28
Deductive reasoning	8	6	11	1	-	18	1.72
Simulation	9	10	6	1	-	17	1.44
Avoid reasoning fallacies	11	8	9	1	-	18	1.61
Judgment	6	13	3	1	1	18	1.44
MONITORING SKILLS							
Knowing when to start critical thinking	7	6	6	5	1	18	2.06
Knowing when to stop critical thinking	9	5	5	7	1	18	2.22
Monitoring progress of critical thinking	2	4	9	4	1	18	2.06
Knowing one's own limitations	6	6	9	3	1	18	1.78
Meta-reasoning	8	10	6	2	-	18	1.61

Participants then were given guided interviews where they were asked to elaborate the answers they had provided on the survey. The sessions, which lasted between 1.5 and 2 hours, concluded with focused interviews with each participant on a wide range of topics pertinent to CT in battle command. These included elaborations on the situational conditions that best typify a CT episode, incidents they had experienced that involved CT, considerations of variables whose presence or absence moderate the effectiveness of CT, as well as discussion of relationships between CT skills and battle command tasks. The interviews concluded with open-ended discussions about the CT skills that participants believed would have highest payoff for battle command as well as areas related to CT where training needs are deemed most pressing.

Results

Participant responses to the written survey were submitted to a variety of analyses, each pertaining to a different component of the CT model. Specifically, analyses of the survey data were conducted to examine the application of CT skills, situation conditions, and predisposing factors to the domain of battle command. The results of the survey data given here are organized by these three model elements.

CT Skills

CT Skills Important To Battle Command. As noted previously, participants were asked to rate the importance of the thirteen CT skill classes to battle command. The officers' responses

on the importance scale were first converted to numerical values using a 4 point scale where 1 represented “very important” and 4 represented “not important at all.” These numerical ratings were then averaged across the eighteen survey respondents for each of the thirteen skills. The mean rating for each class of skill can be seen in Table 5. Table 5 also shows the number of officers that rated each skill as very important, important, somewhat important, and not important, as well as the number who reported each skill class as problematic.

The results showed that each of the 18 officers who participated in the survey had no difficulty understanding the skills and their application to battle command. Of the 234 possible survey responses (18 participants x 13 classes of skills), the officers provided 233 responses, which is a very high response rate (99.6%). These data indicate that the officers could readily comprehend the specific CT skills in question.

The ratings also indicated that all 13 CT skills were considered important, as the mean rating for all classes was lower than 2.3, which is just above the “important” scale anchor. The frequency distribution showed that most skills were rated as either “very important” or “important,” with only 14.9% of responses (35/ 234) falling in the other two categories. Analyses failed to reveal statistically significant differences among the ratings of importance of the 13 categories of CT skills.

CT Skills Problematic for Battle Command. Respondents were also asked to indicate which CT skills were problematic in the conduct of their work. Officers were given space in the survey booklet to check those skills where they had either experienced difficulties themselves or observed others have problems in performing the skill in connection with battle command tasks. The number of participants who reported that they had experienced problems with each skill is also given in Table 5 under the column labeled “Problematic.” Only one of the skill classes, “monitoring progress of critical thinking,” was viewed as problematic by less than one-third of the respondents. In fact, 8 of the 13 skill classes were considered problematic by more than 40% of the sample.

Importance of Situation Conditions to CT in Battle Command. Respondents were asked to evaluate the importance of eleven situation conditions for using CT in battle command. The distribution of importance ratings for each of the eleven situation conditions is shown in Table 6, where the situational conditions are listed in descending order of rated importance. All situational conditions were generally rated as usually or always important. The lowest rated condition, *information-rich stimulus content*, received a mean rating of 2.06, which is near the anchor point of “usually important.” Only five respondents (2.5% of 198 responses) rated any of the conditions as rarely important. Moreover, the officers rated the conditions as “sometimes important” only 24 times of 198 responses (12.1%). Although Table 6 shows that there was some variance in the ratings, the means indicate that all 11 situation conditions were regarded by participants as *usually important* to battle command. Officer comments during the follow-up interviews failed to reveal any additional situation conditions that should be added to the 11 listed below.

Analyses failed to reveal statistically significant differences among the ratings of importance of the 11 situation conditions. However, the ratings indicate that participants believe

that the presence of conflicting information in the stimulus and the consequence of high stakes are generally always important in eliciting a CT episode. They were rated as “always important” by two-thirds of the respondent sample, producing the lowest mean ratings of 1.39 and 1.41, respectively.

It is interesting that the two defining situation conditions posited by the CT model—i.e., those that must be present for CT to occur—were not rated as particularly important to initiating a CT episode. Specifically, the CT model asserts that there must be sufficient time to engage in a CT episode, and that the stimulus material should be substantive or content-rich. As seen in Table 6, these two conditions were rated “always important” only 7 and 4 times, respectively. In

Table 6. Distribution of Importance Ratings for 11 Situation Conditions Associated with CT

Situation Condition or Context	Always Imp. (1)	Usually Imp. (2)	Sometimes Imp. (3)	Rarely Imp. (4)	No. Resp.	Mean Rating
When I see conflicting information in the materials	12	5	1	-	18	1.39
When the stakes are high	12	4	-	1	17	1.41
When there is not a single answer to the problem— multiple answers are possible	10	6	1	1	18	1.61
When there is no preset procedure or algorithm to follow	9	7	1	1	18	1.67
When there is information in the stimulus that is uncertain	6	11	1	-	18	1.72
When both logical and factual information in the materials has to be considered	7	9	2	-	18	1.72
When there is disorder in potentially informative materials	6	8	3	-	17	1.82
When it is important that I fully understand and remember the information in the stimulus	5	10	2	1	18	1.94
When the information in the materials is complex	4	11	3	-	18	1.94
When time is not a limiting factor in reviewing the materials	7	3	7	-	17	2.00
When there is a lot of content-rich information in the materials to be reviewed	4	10	3	1	18	2.06

fact, these were the two lowest rated conditions in the survey, both with means at or above 2.00. However, comments by the respondents suggest that, for the time condition at least, there may have been some confusion as to the intended dimension. Specifically, three participants spontaneously reported that “time is always a limiting factor in battle command,” Thus they were not able to give it a high rating in each case. Had that condition been described instead, as “there is considerable time available,” the resultant ratings might have been much higher, and more in line with those expected of a defining condition. Similarly, battle command situations

are almost always “content-rich” because of the surplus of available information. A different pattern of results may have been observed if the situation condition had been labeled “lack of information”.

The Importance of Predisposing Factors. The distribution of participant ratings of importance for nine predisposing factors is shown in Table 7. Again, these factors are listed in descending order of mean rated importance. Seven of the nine factors were generally regarded as having an important influence on the tendency to use CT in battle command situations, as all had mean ratings below 2.0, the “usually important” anchor point on the scale. Of the 228 responses to this question, only 3 (1.3%) rated a predisposing factor as “rarely important” and only 21 (9.2%) rated a factor as “sometimes important”. Nevertheless, some of the attitudes were rated significantly higher than others, as substantiated by a one-way analysis of variance ($F = 3.544 (8,120), p < .001$). The Army officers who participated in this investigation viewed persistence, reasonableness, confidence, willingness to engage in mental effort, and withholding judgment as the predispositions most strongly associated with a tendency to engage in CT. Interestingly, being skeptical and curious, while identified in the literature as strongly linked to CT, were rated less important. Comments made by participants after the survey did not reveal the presence of any additional noteworthy predispositions.

Table 7. Distribution of Importance Ratings for Nine Predisposing Attitudes Associated with CT

Predisposing Attitude	Always Imp. (1)	Usually Imp. (2)	Sometimes Imp. (3)	Rarely Imp. (4)	No. Resp.	Mean Rating
Person is persistent , is willing to expend effort to tackle a difficult problem	10	7	-	-	17	1.41
Person is reasonable , believes that opinions vary in quality, with good opinions supported by reasons	9	7	2	-	18	1.61
Person has confidence in his/her own reasoning skills	8	8	2	-	18	1.67
Person is willing to engage in extensive mental effort	6	12	-	-	18	1.67
Person does not jump to conclusions but is willing to withhold judgment and gather more information	8	8	2	-	18	1.67
Person is objective , can view things in a detached manner	7	9	2	-	18	1.72
Person does not get anxious when thinking deeply about vital information	4	3	-	1	18	1.89
Person is skeptical , demands justification for assertions	2	9	6	-	17	2.24
Person is intellectually curious , even for information that is not immediately useful or conflicts with own position	3	5	7	2	17	2.47

Discussion

The results of this investigation indicate that the CT model proposed in this report largely captures the skills, situational conditions, and predisposing factors significant to Army battle command. All of the particular instances of these three variables were regarded as, at least, sometimes important to battle command. However, some factors are less important than others, according to our respondents. For example, some of the meta-cognitive classes of skills (e.g., knowing when to start CT) were rated slightly less important. Similarly, skepticism and curiosity were not rated as important to Army battle command as were some of the other predisposing attitudes to CT. Nonetheless, no skill class, situational condition, nor predisposing factor was rated as unimportant by a significant number of respondents.

In contrast, the results of the interviews indicate that only some of the particular instances of skills are problematic to Army battle command. The ability to make judgments, assess the quality of material for consistency, and knowing when to start CT and monitor its progress are *not* skills that are problematic, according to respondents. A pragmatic approach to focusing future research efforts would limit investigation to those skills that are associated with performance deficiencies. Thus, the results of this investigation may be used to identify a set of CT skills that are both important and problematic to Army battle command. The next section of this report describes additional analyses that were conducted on the Fort Hood data to select a set of skills on which to focus future research. The purpose of the analyses was to identify high-payoff skills that are both important and problematic to battle command.

SELECTION OF HIGH-PAYOFF CT SKILLS FOR BATTLE COMMAND

A two-tiered approach was used to select a set of high-payoff CT skills to be trained in (CT)². At the first tier, the field was narrowed by identifying six broad classes of skills from the original 13 that were evaluated by the participants in the Fort Hood interviews. The first cut was made by jointly considering two criteria based on the Fort Hood survey data. First, the mean importance ratings of the 13 broad classes of skills were taken into account. Second, we considered the number of officers who reported that they had observed problems in executing the 13 broad classes of skills.

The mean importance ratings are presented in the far right-hand column of Table 5 and their significance was previously discussed. The number of officers who reported they had observed problems with each skill class is also given in Table 5. Figure 9 plots each CT skill class in terms of its rated importance (x-axis) and number of reported problems (y-axis). To systematize the selection process and equally weight the two criteria, cross-hairs are placed on the figure to demark lines along the x- and y-axes where reasonable cutpoints could be placed. Using this approach, the broad classes of skills located in the upper left quadrant were selected. As shown in Figure 9, there are six such classes that satisfy the criteria of having a sufficiently high importance rating (below 1.7 indicating officers thought they were very important) and a large incidence of problems (more than 7 officers out of 18 reported they had observed problems). Using these criteria, the six classes of CT skills that deserve the most vigorous investigation in future research are:

- Framing the problem
- Inductive reasoning
- Mental simulation
- Avoiding reasoning fallacies
- Meta-reasoning
- Extracting meaning.

However, over 40 CT skills may be associated with these six broad classes. Thus, it was necessary to apply a second cut to narrow the field to a set of particular skills on which to focus further research. We relied heavily upon the officer's incident reports and comments about each class of skill to make the final selection of eight CT skills that would ultimately be the focus of the training system. In addition to participant's comments, we also applied the following four criteria.

- all six classes identified should be represented in the final selection of skills
- the skill should be used in the performance of particular battle command tasks identified as problematic in the Fort Hood survey
- the underlying cognitive processes of the skill should be potentially measurable
- the skill should be applicable to future research and training development efforts.

The greatest weight was given to the second selection criteria. Thus, each of the eight skills was primarily chosen because it was used in a particular battle command task for which deficiencies had been observed by the interviewees.

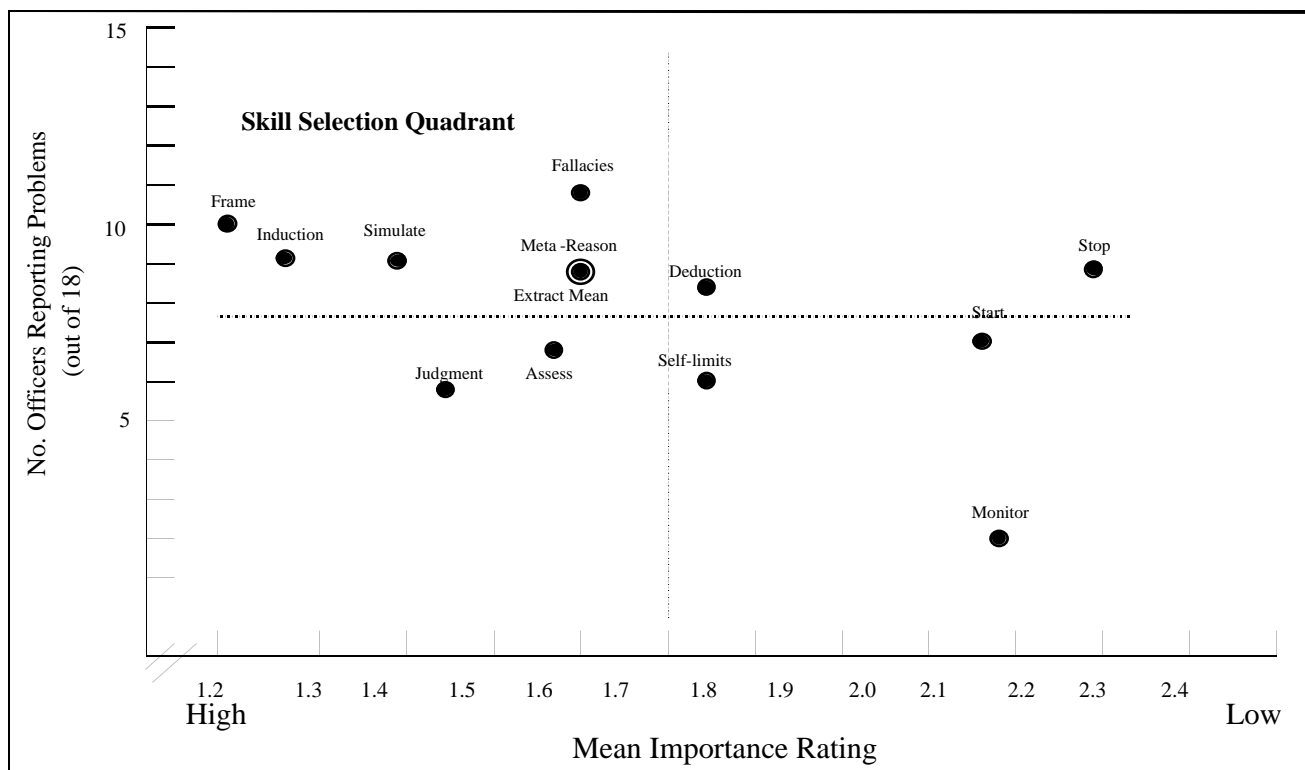


Figure 9. Ratings of Importance (x) by reported problems (y) for 13 broad classes of CT skills.

The greatest weight was given to the second selection criteria. Thus, each of the eight skills was primarily chosen because it was used in a particular battle command task for which deficiencies had been observed by the interviewees.

Application of these criteria resulted in the identification of eight high-payoff CT skills, listed in the middle column of Table 8. The six broad classes of CT skills can be seen in the left-hand column, with associated battle command tasks noted in the right-hand column. These tasks come from a taxonomy of 36 battle command tasks that were consistently found to have serious human performance problems during post and field exercises (Falleesen, 1993). The task list was organized around the taxonomic categories of mission, terrain, enemy, own forces, courses of action (COAs), battlefield, and planning.

The interview data obtained at Fort Hood confirmed that these battle command tasks are problematic, i.e., officers reported that they had observed deficiencies related to CT in these battle-command tasks. Interview respondents identified the battle command tasks for which deficiencies in CT were most evident and troublesome. As noted above, these reports were central to identification and selection of the eight CT skills. Greater weight was given to battle command tasks that were noted by several officers. For example, quite a few of the interviewees reported that commander's intent statements were difficult to write and to understand. The purpose of commander's intent statements is to provide the central objective and intention of the mission. However, that goal is rarely achieved, according to our respondents. Another commonly noted battle command problem was the inability to recognize one's own biases, which manifested itself in "fighting the plan". Many interviewees reported experiences in which new and contrary evidence (e.g., change) was ignored or interpreted to fit the plan. Thus, decisions were made to continue with the original battle plan, often with disastrous results. Table 8 shows the battle command tasks that were reported as problematic in the interviews. The task given in bold is the one that was primarily noteworthy for that CT skill. Secondary problem tasks are listed underneath each primary task. Table 9 further defines and clarifies the problems associated with each of the eight selected CT skills (CTS).

Table 8. Core CT Skills Selected for Training Implementation

CT Skill Class	CT Skill	Associated Battle Command Tasks
Frame the Problem/Extract Meaning from Material	Frame the Message	<ul style="list-style-type: none"> • Clarify intent of commanders 1 and 2 levels up • Review a mission statement • Determine own unit's area of responsibility from OPORD • Determine constraints/restraints placed on mission by higher HQ from OPORD
Extract Meaning from Material	Recognize Gist In Material	<ul style="list-style-type: none"> • Review a mission statement • Clarify commanders' intent 1 and 2 levels up
Induction	Develop An Explanation That Ties Information Elements Together In A Plausible Way	<ul style="list-style-type: none"> • Interpret reports of recent enemy significant activities in area of interest • Clarify intent of commanders one and two levels up • Read the battlefield • Track the battlefield • Interpret reports of enemy disposition
Induction	Generalize From Specific Instances To Broader Classes	<ul style="list-style-type: none"> • Interpret reports of enemy disposition • Interpret reports of recent enemy significant activities in area of interest
Frame the Problem/Simulation	Use Mental Imagery To Evaluate Plans	<ul style="list-style-type: none"> • War game a COA • Develop branches/sequences for each COA • Determine ease of movement • Perform terrain analysis to gain an appropriate perspective that supports battle • Predict number of vehicles that can fit in an engagement area
Avoid Reasoning Fallacies	Challenge One's Bias	<ul style="list-style-type: none"> • Change own-unit plans based on new tactical input • Develop COA's • Read the battlefield • Track the battlefield • Assess the situation • Infer status of enemy forces
Meta-Reasoning	Examine Other Peoples' Perspectives	<ul style="list-style-type: none"> • Interpret reports of recent enemy significant activities in area of interest • Infer status of enemy forces • Interpret reports of enemy disposition • Read the battlefield • Track the battlefield • Assess the situation
Meta-Reasoning	Decide When To Seek Information Based On Its Value And Cost	<ul style="list-style-type: none"> • Assess the situation • Interpret reports of enemy disposition • Interpret reports of recent enemy significant activities in area of interest • Read the battlefield • Track the battlefield • Infer status of enemy forces

Table 9. Relationship of Battle Command (BC) Tasks, Critical Thinking Issues and Selected Critical Thinking Skills (CTS)

CTS #	CTS Title	CTS Definition	Primary BC Task	BC Errors & Deficiencies
1	Frame the Message	The ability to identify the essential elements of a message, understand their relationships, and describe a high fidelity representation of the message.	Clarify the intent of the commanders 1 and 2 levels up	Difficulty in establishing clear and accurate understanding of CDR intent Difficulty in conveying clear CDR intent
2	Recognize Gist in Material	The ability to sort through the details in a message (written, graphical, visual, auditory, and/or tabular) and extract the gist therein.	Restate mission objectives provided by upper echelon to write own mission statement	Too much detail in OPORDS that must be filtered to establish gist that supports writing of own mission statement Too little time at lower echelons to accurately extract essence of mission
3	Develop an Explanation that Ties Information Elements Together in a Plausible Way	The ability to: <ul style="list-style-type: none"> • Arrange evidence logically • Highlight the gaps in knowledge. • Develop an explanation or multiple explanations based on evidence • Evaluate explanation(s) for plausibility 	Interpret reports of recent enemy activities in area of interest to estimate enemy intent and predict enemy actions	Overlook seemingly unrelated facts Fail to assess the quality of information Difficulty in filtering excessive information Tendency to embellish enemy activity reports—over-reports of enemy contact and movement Tendency to discount initial reports
4	Generalize from Specific Instances to Broader Classes.	The ability to recognize and then classify specific facts/incidents/ events as part of a general category.	Interpret reports of enemy disposition	Fail to accurately induce patterns of overall movement based on report instances Tendency to disregard reports that do not match expectations Tendency to inflate information in reports
5	Use Mental Imagery to Evaluate Plans	The ability to accurately create mental images in one's mind about how resources will be applied and events will unfold within a situation.	Develop scheme of maneuver War game COAs	Failure to visualize events Fail to include sufficient detail in COAs Failure to consider contingencies Fail to consider how plans could go wrong Generate only one COA Failure to consider combat multipliers Difficulty in keeping track of mobile forces
6	Challenge One's Bias	The ability to consistently reevaluate one's current view of the situation for prejudice or bias as new information is received.	Change own-unit plans based on new tactical input	Tendency to "fight the plan" General reluctance to change plans

Table 9. (Continued) Relationship of Battle Command (BC) Tasks, Critical Thinking Issues and Selected Critical Thinking Skills (CTS)

CTS #	CTS Title	CTS Definition	Primary BC Task	BC Errors & Deficiencies
7	Examine Other Peoples' Perspectives	The ability to view and interpret a set of circumstances from the perspectives of different individuals, different cultures/religions, and different timeframes (historical perspective).	Interpret reports of recent enemy activities in area of interest	Failure to accurately estimate enemy intent
8	Decide When to Seek Information Based on its Value and Cost.	The ability to evaluate the need for of new information in terms of its cost in: <ul style="list-style-type: none"> • Time • Resources • Risk 	Assess current situation	Tendency to spend too much time planning and gathering information Tendency to make quick decisions without gathering more information

A definition of each skill is given in the third column of Table 9. The primary battle command task for which deficiencies were noted in the interview, and for which the CT skill was chosen, is given in the fourth column. The fifth column describes the kind of errors and deficiencies that were reported by the Fort Hood officers.

CT SKILL TRAINING AT THE COMMAND AND GENERAL STAFF COLLEGE

Background

Army education at all levels is under transformation. Responding to a new world where mission requirements have dramatically changed, the Army has been revising its training curriculum. One transformation has occurred at the Command and General Staff College (CGSC) at Fort Leavenworth, Kansas (Bralley, 2006).

CGSC has developed a new curriculum to meet the changing educational needs of the next generation of Army officers. In place of the Command and Generals Staff Officer's Course (CGSOC), all Army Active component majors attend Intermediate Level Education (ILE). All officers participate in a Common Core Course and then attend an advanced course tailored to their individual specialties. All ILE core courses consist of instruction in four major instructional blocks: Foundations, Strategic Studies, Operational Studies and Tactical Studies. The End of Core Course Exercise forces officers to use critical thinking and critical reasoning skills to analyze and select the best possible courses of action (Bralley, 2006).

In designing the ILE curriculum, one of the focal points of the new curriculum was the requirement to increase students' critical thinking skills. The CT model and CT skills described in this volume was integrated into ILE lesson plans. .

CT Skills Incorporated in Course Curriculum

To assist the ILE design team in developing the ILE curriculum in critical thinking, scientists from the (CT)² project constructed clear labels and definitions of the CT skills. These are shown in Table 9. These definitions are used in the ILE lesson plans and helped to unify the treatment of CT skills among the lessons.

The CT model and skills identified in the research described in this report have been integrated into the ILE team's core course materials. The skills are incorporated into each of the four major blocks of instruction that are being taught in the Common Core course: Foundations, Strategic Studies, Operational Studies, and Tactical Studies. The skills have been integrated into the course modules in the form of lessons plans and/or assessment.

Table 10 identifies the CT skills that were integrated into each of the major blocks of instruction for the Core Course at the beginning of the ILE curriculum design.

Table 10. Critical Thinking Skills Incorporated into the original CGSC-ILE Core Course

Core Course	
Foundations	<ul style="list-style-type: none"> • Generate alternative explanations • Recognize gist in material • Examine other perspectives • Frame the message • Develop and use a mental model • Challenge one's bias
Strategy	<ul style="list-style-type: none"> • Examine other perspectives • Generalize from the specific • Recognize gist in material • Develop and use a mental model • Challenge one's bias • Construct a plausible story • Decide when to seek other info
Operational Art	<ul style="list-style-type: none"> • Construct a plausible story • Visualize plans that meet aims • Examine other perspectives • Frame the message • Recognize gist in material • Decide when to seek other info • Challenge one's bias • Generate alternative explanations
Tactics	<ul style="list-style-type: none"> • Generalize from the specific • Generate alternative explanations • Visualize plans that meet aims • Frame the message • Develop and use a mental model • Construct a plausible story • Recognize gist in material
History	<ul style="list-style-type: none"> • Frame the message • Recognize gist in material • Construct a plausible story • Challenge one's bias • Generate alternative explanations • Examine other perspectives
Leadership	<ul style="list-style-type: none"> • Examine other perspectives • Generalize from the specific • Develop and use a mental model
Force Management	<ul style="list-style-type: none"> • Frame the message • Visualize plans that meet aims
End of Core Course Exercise	<ul style="list-style-type: none"> • Challenge one's bias • Frame the message • Develop and use a mental model

CONCLUSION

The model of CT generated a number of predictions that previously had not been empirically tested. The model was sufficiently specified to permit falsification of many of its assertions, which other models of CT in the literature had not provided. The present investigation tested five of the model's central predictions. As a result of the investigation we now have a clearer picture of the effects of judgment and understanding tasks on CT and the effects of stimulus substance on CT. It is now clear that CT does not always generate negative affect and that experience may well increase CT.

Although the results of the validation experiment were mixed in their support of the model, the model has passed an important scientific criterion. It has generated testable hypotheses that have produced empirical findings from which we have gained knowledge. Some of the findings point to places in the model that require greater specification or modification. Other findings are consistent with the model's predictions.

These results also have practical implications for the design of information systems and for educational and training programs that seek to increase the use of CT skills. Designers and teachers should be aware that people may not question highly substantive material any more than low substantive material. If CT is desired, inconsistent content might be highlighted by information systems. Similarly, if educational and training programs seek to encourage CT, one strategy would be to sensitize students to inconsistent material.

The CT skills identified in this research as problematic to Army battle command may be utilized for training and assessment purposes and to increase self-awareness. These skills have now been integrated into the Army Command and General Staff College Intermediate Level Education (ILE) course materials. The original skills identified in the first phase of the present research are now incorporated into each of the five major blocks that are being taught in the ILE Common Core course. These skills have been integrated into 16 modules in the form of lessons plans and/or assessment. The Common Core course incorporated the skills into 63 lesson plans. Also, training concepts discussed in Fischer, Spiker and Riedel (2008b) have been adopted to teach the skills in ILE.

REFERENCES

- American Philosophical Association. (1990). *The Delphi Report. Committee on Pre-College Philosophy* (ERIC Doc. No. ED 315 423.)
- Baker, P. J., & Anderson, L. E. (1987). *Social Problems: A critical thinking approach*. Belmont, CA: Wadsworth.
- Baron, J. B. & Sternberg, R. J. (1986). *Teaching Thinking Skills*. New York, NY: W. H. Freeman & Company.
- Beyer, B. K. (1995). *Critical thinking*. Bloomington, IN: Phi Delta Kappa Educational Foundation.
- Bralley, N.H. (2006). ILE: A new system for CGSC students. *Army Logistician*. 38 (1). Retrieved from www-cgsc.army.mil/carl/download/csipubs/infantry/inf_intro_cvii.pdf.
- Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, 48, 306-307.
- Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39, 752-766.
- Chaiken, S., & Trope, Y. (1999). *Dual-process theories in social psychology*. New York: Guilford Press.
- Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic processing within and beyond the persuasion context. In J. S. Uleman & J. A. Bargh (Eds.). *Unintended thought*. New York: Guilford Press.
- Cohen, M. S., Adelman, L., Tolcott, M.A., Bresnick, T. A., Marvin, F. F. (1994). *A cognitive framework for battlefield commanders' situation assessment*. ARI Technical Report 1002. Alexandria, VA: U.S. Army Research Institute.
- Cohen, M. S., Freeman, J. T., Fallesen, J. J., Marvin, F. M., & Bresnick, T. A. (1996). *Training critical thinking skills for battlefield situation assessment: An Experimental test*. ARI Technical Report 1050. Alexandria, VA: U.S. Army Research Institute.
- Cohen, M. S., Thompson, B. B., Adelman, L., Bresnick, T. A., & Riedel, S. L. (1999). *Training battlefield critical thinking and initiative*. (Interim Report). Cognitive Technologies, Inc.: Alexandria, VA. .
- D'Angelo, E. (1971). *The teaching of critical thinking*. N.V. Amsterdam: B.R. Grüner.

- Department of the Army. (2001). Field Manual 3-0, *Operations*. Washington, DC: Headquarters, Author. Retrieved 13 June 2008, from <http://www.globalsecurity.org/military/library/policy/army/fm/3-0/ch5.htm>.
- Ennis, R. H. (1962). A concept of critical thinking. *Harvard Educational Review*, 32 (1), 81-111.
- Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron & R. J. Sternberg (Eds.), *Teaching thinking skills: Theory and practice*. New York, NY: W. H. Freeman.
- Ennis, R. H. (1996). Critical thinking: what is it? ©1996 *Philosophy of Education Society*, from the World Wide Web: http://www.ed.uiuc.edu/PES/92_docs/Ennis.HTM
- Ennis, R. H., Millman, J., & Tomko, T. N. (1985). *Cornell critical thinking tests level X & Level Z*. (3rd ed.) Pacific Grove, CA: Midwest Publications.
- Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. London: Erlbaum Associates.
- Evans, J. St. B. T. & Over, D. E. (1996). *Rationality and reasoning*. Hove, England: Psychology Press.
- Facione, N. C. (1995). *Critical thinking & clinical Judgment: goals 2000 for nursing science*. Paper presented at the annual meeting of the Western Institute of Nursing. San Diego, CA.
- Facione, N. C., Facione, P. A., & Sanchez, C. A. (1994). Critical thinking disposition as a measure of competent clinical Judgment: The development of the California Critical Thinking Disposition Inventory. *Journal of Nursing Education*, 33, 345-350.
- Facione, P. A., Facione, N. C., Blohm, S.W., Howard, K., & Giancarlo, C. A. F. (1998) *California critical thinking skills test*. (Rev. ed.). Millbrae, CA: California Academic Press.
- Fallesen, J. J., Michel, R. R., Lussier, J. W., & Pounds. J. (1996). *Practical thinking: innovation in battle command instruction*. Technical Report 1037. Alexandria, VA: U.S. Army Research Institute
- Farley, M. J., & Elmore, P. B. (1992). The relationship of reading comprehension to critical thinking skills, cognitive ability, and vocabulary for a sample of underachieving college freshmen. *Educational and Psychological Measurement*, 52, 921-931.
- Fischer, S.C. Spiker, V. A., & Riedel, S.L. (2008a). *Critical thinking training for Army officers. Volume One: Overview of the research program*. ARI Research Report 1881. Arlington, VA: US Army Research for the Behavioral and Social Sciences.
- Fischer, S.C. Spiker, V. A., & Riedel, S.L. (2008b). *Training critical thinking for Army officers. Volume Three: Development and assessment of a web-based program*. ARI Research Report 1883. Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

- Fischer, S.C. Spiker, V. A., Harris, D.H., McPeters, E.V., & Riedel, S.L. (2008). *Computerized training in critical thinking (CT)²: A skill-based program for Army personnel*. ARI Research Report 1880. Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Frisby, C. L. (1991). A meta-analytic investigation of the relationship between grade level and mean scores on the Cornell Critical Thinking Test (Level X). *Measurement and Evaluation in Counseling and Development*, 23, 162-170.
- Frisby, C. L. (1992). Construct validity and psychometric properties of the Cornell Critical Thinking Test (Level Z): A Contrasted Groups Analysis. *Psychological Reports*, 71, 291-303.
- Gadzella, B., Ginther, D. W., & Bryant, G. W. (1997). Prediction of performance in an academic course by scores on measures of learning style and critical thinking. *Psychological Reports*, 81, 595-602.
- Gadzella, B., & Masten, W. G. (1998a). Critical thinking and learning processes for students in two major fields. *Journal of Instructional Psychology*, 25(4) 256-261.
- Gadzella, B., & Masten, W. G. (1998b). Relation between measures of critical thinking and learning styles. *Psychological Reports*, 83, 1248-1250.
- Gadzella, B., & Penland, E. (1995). Is creativity related to scores on critical thinking? *Psychological Reports*, 77, 817-818.
- Glaser, E.M. (1941). *An experiment in the development of critical thinking*. New York, NY: Teachers College Columbia University.
- Gonzalez, E. W. (1996). Relationships of nurses' critical thinking ability and perceived patient self-disclosure to accuracy in assessment of depression. *Issues in Mental Health Nursing*, 17, 111-122.
- Gubbins, E. J. (1986). Matrix of Thinking Skills. In R. J. Sternberg, *Critical Thinking: Its Nature, Measurement and Improvement*. New Haven, CT: Yale University.
- Halpern, D. F. (1992). Notes on enhancing thinking skills in the sciences and mathematics. In D. F. Halpern (Ed.), *Enhancing Thinking Skills in the Sciences and Mathematics*. Hillsdale, NJ: LEA.
- Halpern, D. F. (1996). *Thought and knowledge: an introduction to critical thinking* (3rd Ed.). Mahwah, NJ: L. Erlbaum Associates.
- Halpern, D. F. (1997). *Critical thinking across the curriculum: A brief edition of thought and knowledge*. Mahwah, NJ: Lawrence Erlbaum.

- Halpern, D. F. (1998). Teaching critical thinking for transfer across domains. *American Psychologist*, 53(4), 449-455.
- Hawley, D. A. (1998). *The measurement of a critical thinking disposition among practicing registered nurses*. Doctoral Dissertation, New Mexico State University.
- Holyoak, K. J. & Spellman, B. A. (1993). Thinking. *Annual Review of Psychology*, 44, 265-316.
- Infantry in Battle*, (1939). Washington, D.C.: The Infantry Journal Incorporated. Retrieved from www-cgsc.army.mil/carl/download/csipubs/infantry/inf_intro_cvii.pdf
- Jacobs, S. (1999.) The equivalence of forms A and B of the California Critical Thinking Skills Test. *Measurement and Evaluation in Counseling and Development*, 31, 211-222.
- Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.
- Johnson-Laird, P. N. (1993). How the mind thinks. In G. Harman, (Ed), *Conceptions of the human mind; Essays in honor of George A. Miller*. Hillsdale, NJ, Lawrence Erlbaum Associates.
- Kahneman, D. (2000). A psychological point of view: Violations of rational rules as a diagnostic of mental processes. (Commentary on Stanovich and West). *Behavioral and Brain Sciences*, 23, 681-683.
- Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58, 697-720.
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.) *Heuristics and biases*. Cambridge, U.K.: Cambridge University Press.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, U.K.: Cambridge University press.
- Keeley, S. M. (1992). Are college students learning the critical thinking skill of finding assumptions? *College Student Journal*, 26, 316-322.
- Klein, G. (1999). *Sources of power*. Cambridge, MA: MIT Press.
- Klein, G. A., Orasanu, J., Calderwood, R., & Zsombok, C. E. (1993). *Decision-making in action: Methods and models*. Norwood, NJ: Ablex.
- Kurfiss, J. G. (1988). *Critical thinking: theory, research, practice, and possibilities*. Washington, D.C.: Association for the Investigation of Higher Education.

- Lauer, R. (1998). Overview. In S. P. Kodish & R. P. Holston (Eds.), *Developing sanity in human affairs*. Westport, CT: Greenwood Press.
- Lawson, T. J. (1999). Assessing psychological critical thinking as a learning outcome for psychological majors. *Teaching of Psychology*, 26, 207-209.
- Massaro, D. W. (1997). What rethinking reason requires. *American Journal of Psychology*, 110(2), 285-314.
- McBride, R. E., & Bonnette, R. (1995). Teacher and at-risk students' cognitions during open-ended activities: Structuring the learning environment for critical thinking. *Teaching and Teacher Education*, 11, 373-378.
- McBride, R. E., Reed, J. (1998). Thinking and college athletes – are they predisposed to critical thinking? *College Student Journal*, 443-451.
- McCarthy, C. (1996). Why be critical? (or rational, or moral?) on the justification of critical thinking. ©1996 *Philosophy of Education Society*, from the World Wide Web: http://www.ed.uiuc.edu/PES/92_docs/McCarthy.HTM
- McCutcheon, L. E., Apperson, J. M., Hanson, E., Wynn, V. (1992). Relationships among critical thinking skills, academic achievement, and misconceptions about psychology. *Psychological Reports*, 71, 635-639.
- McPeck, J. (1996). Underlying traits of critical thinkers: A response to Stephen Norris. (1996) *Philosophy of Education Society*, from the World Wide Web: http://www.ed.uiuc.edu/PES/92_docs/McPeck_to_Norris.HTM
- Meehl, P. E. (1950). On the circularity of the law of effect. *Psychological Bulletin*, 47, 52-75.
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minneapolis Press.
- Miller, M. A., & Malcolm, S. (1990). Critical thinking in the nursing curriculum. *Nursing and Health Care*, 11(4), 67-73.
- Moore, B. N. & Parker, R. (1989). *Critical thinking: Evaluating claims and arguments in everyday life*. Mountain View, CA: Mayfield.
- Moore, W. G., McCann, H., McCann, J. (1985). *Creative and critical thinking* (2nd ed.). Boston, MA: Houghton Mifflin.
- Moss, P.A., & Koziol, S.M. (1991). Investigating the validity of a locally developed critical thinking test. *Educational Measurement: Issues and Practice*, 17-22.

- National Education Goals Panel. (1991). The national education goals report. Washington, DC: U.S. Government Printing Office.
- Norris, S. P. (1985). Synthesis of research on critical thinking. *Educational Leadership*, 42, 40-45.
- Norris, S. P. (1989). Can we test validly for critical thinking? *Educational Researcher*, 18, 21-26.
- Norris, S. P. (1996) Bachelors, buckyballs, and ganders: Seeking analogues for definitions of "critical thinker." ©1996 *Philosophy of Education Society*, from the World Wide Web: http://www.ed.uiuc.edu/PES/92_docs/Norris.HTM
- Norris, S. P. & Ennis, R. H. (1989). *Evaluating critical thinking*. Pacific Grove, CA: Critical Thinking Press & Software.
- Paul, R. W. (1995). *Critical thinking*. Santa Rosa, CA: Foundation for Critical Thinking.
- Paul, R. & Elder, L. (2002). *Critical thinking – Tools for taking charge of your learning and your life*. Up Saddle River, NJ: Prentice Hall.
- Pellegrino, J. W. (1995). Technology in support of critical thinking. *Teaching of Psychology*, 22(1), 11-12.
- Perkins, D., Jay, E., & Tishman, S. (1993). Beyond abilities: A dispositional theory of thinking. *Merrill-Palmer Quarterly*, 39(1), 1-21.
- Perkins, R. M. (1993). Personality variables and implications for critical thinking. *College Student Journal*, 106-111.
- Perkins, R. J. M. (1993). Personality variable and implications for critical thinking. *College Student Journal*, 106-111.
- Petty, R., & Caccioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer Verlag.
- Posner, M. I., & Snyder, C. R. R. (1975). Attention and cognitive control. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium*, Hillsdale, NJ: Erlbaum.
- Quellmalz, E. S. (1987). Developing reasoning skills. In J.B. Baron & R.J. Sternberg (Eds.), *Teaching thinking skills: Theory and practice*. New York, NY: W.H. Freeman.
- Rasmussen, J., Pejtersen, A., & Goodstein, L. (1992). *Cognitive engineering concepts and applications, Part I: Concepts*. New York, Wiley.

- Resnick, L. B. (1987). *Education and learning to think*. Washington, D.C.: National Academy Press.
- Royalty, J. (1994). Undergraduates' class standing and critical thinking. *Psychological Reports*, 75, 1402.
- Royalty, J. (1995). The generalizability of critical thinking: paranormal beliefs versus statistical reasoning. *The Journal of Genetic Psychology*, 156(4), 477-488.
- Sa, W. C., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, 91, 497-510.
- Schaub, L. J. (1991, September). *The development and retention of critical thinking dispositions among students of the Air Force Institute of Technology graduate management programs* (ADA246669). Wright Patterson AFB, OH: Air Force Institute of Technology.
- Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, 84 (1), 1-66.
- Siegel, H. (1988). *Educating reason*. London: Routledge, Inc.
- Siegel, H. (1996). On defining "critical thinker" and justifying critical thinking. ©1996 *Philosophy of Education Society*, from the World Wide Web: http://www.ed.uiuc.edu/PES/92_docs/Siegel.HTM
- Simon, H. (1957) *Models of man: Social and rational*. New York: Wiley.
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3-22.
- Sloman, S. A. (2002). Two systems of reasoning. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases*. Cambridge, U.K.: Cambridge University Press.
- Stanovich, K. E. & West, R. F. (2000) Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23, 645-726.
- Sternberg, R. J. (1987). Teaching intelligence: The application of cognitive psychology to the improvement of intellectual skills. In J. B. Baron & R. J. Sternberg (Eds.), *Teaching thinking skills: Theory and practice*. New York, NY: W. H. Freeman.
- Swartz, R. J. (1998). Thinking critically about sources of information. In S. P. Kodish & R. P. Holston (Eds.), *Developing sanity in human affairs*. Westport, CT: Greenwood Press.
- Tucker, R. W. (1996). *The editor's desk: Less than critical thinking*. The Phoenix Institute, VI (4).

- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76, 105-110.
- Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5, 207-232.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211, 453-458.
- Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293-315.
- Walsch, C.M., & Hardy, R.C. (1997). Factor structure stability of the California Critical Thinking Disposition Inventory across sex and various students' majors. *Journal of Perceptual and Motor Skills*, 85(3, pt. 2), 1211-1228.
- Walsch, C. M., & Hardy, R. C. (1999). Dispositional differences in critical thinking related to gender and academic major. *Journal of Nursing Education*, 38(4), 149-155.
- Walters, K. S. (Ed.). (1994). *Re-thinking reason: New perspectives in critical thinking*. Albany: State University of New York Press.
- Watson, G., & Glaser, E. M. (1980). *Critical Thinking Appraisal*. New York, NY: Psychological Corporation.
- Youssef, F.A., & Goodrich, N. (1996). Accelerated versus traditional nursing students: A comparison of stress, critical thinking ability and performance. *International Journal of Nursing Studies*, 33, 76-82.
- Zsombok, C.E. (1997). Naturalistic decision-making. In C.E. Zsombok and G. Klein (Eds.), *Naturalistic decision-making*. Mahwah, NJ: Erlbaum, 3-16.

APPENDIX A: CRITICAL THINKING SKILLS EXTRACTED FROM LITERATURE

Interpretation Skills

1. Recognize gist in material
2. Break goal into sub-goals
3. Strip verbal argument of irrelevancies and rephrase it in essential terms
4. Extract meaning from context
5. Understand contextual nuances
6. Frame the message
7. Probe question or problem to obtain clarifying information
8. Question deeply
9. Redefine problem and goal
10. Seek clear statement of the question
11. Understand intended definition of certain words
12. Discern when a term is used with different meanings
13. Recognize need for operational definition
14. Identify and challenge assumptions
15. Identify unstated assumption in a discussion
16. Identify missing information from an argument
17. Identify premises and conclusions
18. Identify missing premises
19. Analyze ambiguities in arguments
20. Critique to distinguish reliable from unreliable assumptions
21. Distinguish fact/opinion/assumption elements in an argument
22. Distinguish relevant from irrelevant information
23. Examine evidence to distinguish anecdote from fact
24. Determine whether a statement is overly vague or overly specific
25. Identify own assumptions and biases
26. Identify emotional language in external sources
27. Distinguish between validity of a belief and intensity with which it is held
28. Seek disconfirming evidence
29. Know when causal claims can and can't be made

Reasoning Skills

1. Understand limits of extrapolation
2. Reason by finding analogous arguments to bolster conclusion
3. Refine generalizations and avoid oversimplification
4. Apply general principles to specific cases
5. Generalize from specific instances to broader classes
6. Determine whether a simple generalization is warranted
7. Draw inductive inference from observations
8. Reason by taking representative samples
9. Distinguish between deductive and inductive reasoning
10. Reason by deductive logic to draw conclusions from premises
11. Reason dialogically to identify and compare perspectives
12. Reason dialectically to evaluate points of view
13. Trace logic in an argument
14. Determine whether a statement follows from premises
15. Distinguish between logically valid and invalid inferences
16. Check consistency of information in the problem
17. Avoid ad hominem reasoning fallacy (consider argument not the person)
18. Avoid false dichotomy reasoning fallacy (artificially reduce the number of choices)
19. Avoid guilt by association reasoning fallacy
20. Avoid emotional appeal reasoning fallacy
21. Identify instances of faulty thinking
22. Mentally simulate plans to see if they achieve goals
23. Mentally generate a structure of possibilities that presently don't exist
24. Mentally simulate probable consequences of alternative
25. Develop and use mental models
26. Recognize bias in hindsight analysis
27. Reason from starting point with which one disagrees
28. Recognize fallibility of own opinion
29. Recognize probability of bias in own opinion
30. Recognize transitive relationships
31. Develop an explanation that ties information elements together in a plausible way
32. Make reasoned value judgment by considering background, consequences, principles
33. Understand difference between reasoning and rationalizing
34. Analyze problem by working backward
35. Determine when evidence is insufficient to warrant sound conclusion

Assessment Skills

1. Know value and cost of information, how and when to seek it
2. Know when new information supports/refutes conclusion
3. Consider new evidence as it becomes available
4. Weigh multiple factors when necessary
5. Perform means-ends analysis to check status
6. Support general assertions with details
7. Frame decision in alternative ways
8. Assess an assertion's truthfulness based on accuracy of relevant facts
9. Assess an assertion's truthfulness based on its degree of precision
10. Assess an assertion's truthfulness based on presence of unbiased evidence
11. Assess an assertion's truthfulness based on having credible sources
12. Assess an assertion's truthfulness based on its logical consistency
13. Assess an observation's credibility based on short time between observation and report
14. Assess an observation's credibility based on first-hand report by observer
15. Assess an observation's credibility based on minimal interference
16. Assess an observation's credibility based on reporter's belief that observation was accurate
17. Assess an observation's credibility based on corroboration by other sources
18. Assess credibility of information source based on author's reputation for accuracy
19. Assess credibility of information source based on being in author's field of expertise
20. Assess credibility of information source based on absence of conflict of interest
21. Assess credibility of information source based on known risk to author's reputation
22. Assess credibility of information source based on data-gathering and processing methods
23. Assess credibility of information source based on agreement with other sources
24. Assess strength of conclusion based on reasonableness of assumptions
25. Assess strength of conclusion based on consistency with known facts
26. Assess strength of conclusion based on alternatives are inconsistent with known facts
27. Assess strength of conclusion based on its ability to explain the evidence
28. Assess strength of argument based on clarity of meaning
29. Assess strength of argument based on identity of stated and unstated conclusions
30. Assess strength of argument based on identity of premises supporting conclusions
31. Assess strength of argument based on identity of unstated assumptions
32. Assess strength of argument based on reliability and reasonableness of inferences
33. Assess strength of argument based on other relevant arguments
34. Assess quality of judgment based on kind of judgment being made
35. Assess quality of judgment based on presence of indicators related to the criteria
36. Assess quality of judgment based on pattern among indicators
37. Assess quality of judgment based on degree of match with criteria

Meta-Cognitive Skills

1. Look beyond first obvious explanation to consider alternative interpretations
2. Identify the need to think hard
3. Develop perspective to explore the implications of beliefs, arguments, or theories
4. Ask questions and be willing to ponder (e.g., use scientific method)
5. Generate summaries
6. Generate alternative explanations
7. Generate multiple ideas
8. Adopt multiple perspectives
9. Consider multiple sides of an issue
10. Stay relevant to the main point
11. Take total situation into account
12. Monitor events for consistency with expectations
13. Monitor own understanding of problem
14. Compare analogous situations and transfer pertinent learning to new contexts
15. Know that observations are more credible than inferences based on them
16. Read between the lines
17. Determine whether argument depends on an ambiguity
18. Understand differences among conclusions, assumptions, and hypotheses
19. Use mental imagery to evaluate plans
20. Challenge one's bias
21. Examine other peoples' perspective
22. Decide when to seek information based on its value and cost

APPENDIX B: PREDISPOSING FACTORS FOR CRITICAL THINKING

1. Knows that opinions vary in quality, with good opinions supported by reasons
2. Intellectual honesty
3. Skeptical
4. Fair-minded
5. Respects clarity and precision
6. Demands justification
7. Reflective cognitive style
8. Persistence (some people opt not to begin the thinking process)
9. Ability to “break set”
10. Having incremental (as opposed to an entity) view of intelligence
11. Willing to suspend judgment and gather more information
12. Aware of own gaps in knowledge
13. Concern for accuracy
14. Maintains an open attitude
15. Adaptability
16. Objectivity
17. Cognitive flexibility to detach reasoning from prior knowledge
18. Desire for knowledge even if it undermines own cause
19. Open-mindedness—tolerance for other views
20. Analyticity—anticipate consequences
21. Confidence in own reasoning skills
22. Intellectual curiosity—eagerness to learn even when knowledge is not immediately useful
23. Cautious in making judgments and aware that more than one solution may be acceptable
24. Willing to work cooperatively
25. Willing to listen to other ideas
26. Disposed to seek reasons
27. Disposed to be well informed
28. Disposed to consider outside points of view
29. Disposed to seek as much precision as possible

APPENDIX C: INSTRUCTIONS USED TO MANIPULATE TASK IN VALIDATION INVESTIGATION

Understanding:

Imagine that you are a psychology student who is trying to explain the study described in the following paragraph. Your task is to orally demonstrate your understanding of the information it contains by talking through your thoughts as you try to comprehend the paragraph.

Judging:

Imagine you are an editor for a research journal who has to decide whether or not you will consider publishing the study described in the following paragraph. Read through the paragraph and talk through your thoughts as you try to decide whether to accept or reject the study.

General Topic:

Imagine that you are a research assistant whose task is to sort through paragraphs describing studies and label each one according to the general topic of the paragraph. Read the following paragraph and talk through your decision of the general topic.

APPENDIX D: TWENTY-SEVEN PARAGRAPHS USED IN VALIDATION INVESTIGATION

Topic: Effects of Caffeine on Memorization

Low Substance:

Researchers have long been interested in the effects of the popular stimulant caffeine on memory. Since the majority of Americans take caffeine every day in their morning coffee, or in their soft drinks, many would like to know whether it improves their memory, as hoped, or has no effect or makes memory worse. To answer this question, a study was conducted to examine the effect of caffeine on performance in a memorization task. The researchers compared performance between subjects who had taken a dose of caffeine before completing the memorization task and those who had not. It was found that the ingestion of the caffeine did indeed result in improved performance in the memorization task, much to the delight of coffee-drinkers everywhere.

Consistent High Substance:

Some researchers wanted to see how caffeine would affect performance in a memorization task. They conducted a study in which fifty college students were randomly assigned to two groups of 25. One group received a dose of caffeine prior to the experiment, while the other received a placebo. Twenty minutes after dosing, each subject was required to learn a list of 15 nonsense syllables presented one at a time. The subject viewed the list once, and then, on each subsequent presentation of the list, tried to predict each syllable before it was shown. The group receiving the caffeine required fewer trips through the list ($M = 14.3$) to reproduce it without mistakes than did the placebo group ($M = 16.8$), $t(23) = 2.36$, $p < .05$, two-tailed.

Inconsistent High Substance:

Some researchers wanted to see how caffeine would affect performance in a memorization task. They conducted a study in which fifty college students were divided into two groups: those who drank coffee regularly and those who did not. One group received a dose of caffeine prior to the experiment, while the other received a placebo. Two hours after dosing, each subject was required to learn a list of 15 nonsense syllables presented one at a time. The subject viewed the list once, and then, on each subsequent presentation of the list, tried to predict each syllable before it was shown. The group receiving the caffeine required fewer trips through the list ($M = 14.3$) to reproduce it without mistakes than did the placebo group ($M = 16.8$), $t(23) = 2.36$, $p < .05$, two-tailed.

Topic: Numerosity in Children

Low Substance:

It is generally assumed that children who have not yet learned to speak are not able to count, but do they have some understanding of the concept of quantity? Can a child recognize that a picture of two cows is different from a picture of one cow? Or that two chickens are different from one chicken? This is the question that researchers attempted to answer in a recent study on young children and their perception of quantity. The researchers studied very young children who had not yet learned to speak to see if they could tell the difference between pictures depicting one object and pictures depicting two objects. By measuring how long a child looked at a newly presented picture, they could tell whether the child thought of the new picture as different from the previous picture. After showing many pictures of one object, such as one cow, one chicken, etc., they would then show a picture of two cows or two chickens, and see if the child looked at the new picture for a longer time than the previous pictures. In this way, they were able to determine that children are in fact able to tell the difference between one and two objects.

Consistent High Substance:

A study was conducted to determine if very young children are able to discriminate between visual displays containing different numbers of items. 96 children between the ages of 10 and 12 months (not yet speaking) were shown pictures, one at a time, of several items varying in size, type, and placement, with only the number of items remaining the same (2, 3, or 4), until the length of time spent looking at each picture had reduced to a consistently short period. When that happened, it was assumed that they recognized a sameness between the various displays. They were then shown several test trials in which the number of items either was the same (habituated display) or differed by one (test display). If the child spent significantly more time looking at the test displays than at the habituated displays, it was assumed the child had recognized the difference in number between them. Results indicated that the children were able to discriminate between 2 and 3 objects, $F(1, 28) = 12.78, p < .001$, and between 3 and 4 objects, $F(1, 28) = 8.92, p < .006$, but not between 4 and 5 objects.

Inconsistent High Substance:

A study was conducted to determine if very young children are able to discriminate between visual displays containing different numbers of items. Some children between the ages of 10 and 12 months (not yet speaking) were shown pictures, one at a time, of several items varying in size, type, and placement, with only the number of items held constant (2, 4, or 6), until the length of time spent looking at each picture exceeded a predetermined limit. When that happened, it was assumed that they recognized the difference between the various displays. They were then shown several test trials in which the number of items either was the same (type display) or differed by one (placement display). If the child spent significantly more time looking at the placement displays than at the type displays, it was assumed the child had not recognized the difference in number between them. Results indicated that the children were able to discriminate

between 2 and 3 objects, $F(1, 28) = 12.78, p < .001$, and between 4 and 5 objects, $F(1, 28) = 8.92, p < .006$, but not between 3 and 4 objects.

Topic: Effects of Day-Care

Low Substance:

Recently a lot of attention in the media has been focused on how time spent in day care facilities affects children. Educators and parents alike are concerned that preschool-aged children who spend more time in day care may develop behavioral problems. In particular, they are concerned about the possibility of day care resulting in an increase in aggressive behaviors in the children. In an attempt to answer this question, a study was done to look at how children are affected by time spent in day care. The study compared the aggressive behaviors of preschool-aged children who attended day care to the aggressive behaviors of those who did not attend day care. The researchers did not find any significant relationship between time spent in day care and aggressive behavior. Thus, the study did not support the hypothesis that children who spend more time in day care will become more aggressive than those who spend less time.

Consistent High Substance:

A research study was conducted to look at the relationship between aggressive behaviors in children and their time spent in day care facilities. Forty-six preschool-aged children were randomly chosen from three day-care facilities and divided into two groups: those who attended half-time ($M=20.2$ hrs/week) and those who attended full-time ($M=38.4$ hrs/week). The children were observed, one at a time, interacting with unfamiliar children in a laboratory setting, and the instances and strengths of three aggressive behaviors were coded and quantified: Assertiveness, Persistence, and Combativeness. Results showed the full-time children exhibiting significantly greater Assertiveness and Persistence than the half-time children, $t(44) = 2.83, p < .01$, two-tailed, and $t(44) = 2.96, p < .01$, two-tailed, respectively. The difference on the third measure, Combativeness, was not significant.

Inconsistent High Substance:

The American Council on Family Morals conducted a research study to look at the relationship between aggressive behaviors in children and their time spent in day care facilities. Ten preschool-aged children were randomly chosen from some day-care facilities and divided into two groups: those who attended half-time ($M=34.5$ hrs/week) and those who attended full-time ($M=38.4$ hrs/week). The children were observed, one at a time, interacting in a laboratory setting, and the instances and strengths of three aggressive behaviors were coded and quantified: Assertiveness, Persistence, and Combativeness. Results showed the full-time children exhibiting significantly greater Assertiveness and Persistence than the half-time children, $t(44) = 2.83, p < .01$, two-tailed, and $t(44) = 2.96, p < .01$, two-tailed, respectively. The difference on the third measure, Combativeness, was not significant. The researchers concluded that children in day care facilities were destined to become the bullies of the playground.

Topic: Effects of Girls' Sports Program

Low Substance:

The governor of a particular state was interested in the adequacy of girls' sports programs in the state's public schools. She was concerned that the existing sports programs for girls may not be adequately meeting the demand for such programs. To determine whether the public schools had adequate sports programs for girls, the governor asked her advisors to research the issue. The advisors reviewed the most recent and accurate research to see if girls' sports programs were adequate. According to their research, more girls were interested in participating in sports programs than had access to them, suggesting that existing programs did not adequately meet demand.

Consistent High Substance:

The governor of a Southern state wanted to determine if more money should be allocated to girls' sports in the state's public schools. To this end, her advisors reviewed a 2002 survey in which 1,256 high school girls in 7 Southern states were chosen by stratified random sampling and polled about their interest in and access to sports programs. Of the 162 girls interviewed in the governor's state, 42 (26%) were interested in participating in school sports but unable to do so because of a lack of programs or inadequate existing programs. This percentage was compared to the average for the remaining 6 states (18%) and was found to be significantly higher, $X^2(1, n=1254) = 8.45, p < .005$.

Inconsistent High Substance:

The governor of a Southern state wanted to determine if more money should be allocated to girls' sports in the state's public schools. To this end, her advisors reviewed a 1954 survey in which 1,256 high school girls in two states were chosen by stratified random sampling and polled about their interest in and access to sports programs. Of the 162 girls interviewed in the governor's state, 42 (26%) were interested in participating in school sports but unable to do so because of a lack of programs or inadequate existing programs. This percentage was compared to the average for the remaining state (18%) and was found to be significantly higher, $X^2(1, n=1254) = 8.45, p < .005$.

Topic: Effects of Hormone Replacement Therapy (HRT)

Low Substance:

Recently there has been a controversy over the health benefits of hormone replacement therapy (HRT) for postmenopausal women. Previously, it had been thought that HRT held the potential for significant health benefits for postmenopausal women, including decreased risks for several common diseases. Many physicians were routinely prescribing HRT to their postmenopausal patients for these supposed benefits. But recently an extensive study was done to examine the actual health benefits of HRT in a large group of postmenopausal women, and the results have

surprisingly shown more risks and fewer benefits associated with HRT than had been previously expected. These findings have led many physicians to reconsider prescribing HRT to their postmenopausal patients.

Consistent High Substance:

A recent study has re-evaluated the health effects of hormone replacement therapy (HRT) for postmenopausal women. The study followed 16,608 healthy, postmenopausal women, age 50-79, for five years. The women were randomly assigned to receive either HRT (estrogen and progesterone) or placebo for the duration of the study. The study's main goal was to see if HRT helped reduce the incidence of heart disease and hip fractures, as suggested by previous studies, and if those potential benefits outweighed the known risks for breast cancer, uterine cancer, and blood clots. The incidences of these diseases were followed and compared between the two groups. While the HRT group did experience significantly fewer hip fractures, an unexpected significant increase was seen in the incidence of heart disease, $X^2(1, n=16608) = 83.6, p<.001$, in addition to the expected increases in breast cancer, uterine cancer, and blood clots.

Inconsistent High Substance:

A recent study has re-evaluated the health effects of hormone replacement therapy (HRT) for postmenopausal women. The study followed 16,608 healthy women, age 31-49, for five years. The women were randomly assigned to receive one of three treatments, HRT, pHRT, or placebo for the duration of the study. The study's main goal was to see if pHRT helped reduce the incidence of heart disease and hip fractures, as suggested by previous studies, and if those potential benefits outweighed the known increased risks for breast cancer, uterine cancer, and blood clots associated with HRT. The incidences of these diseases were followed and compared between the two groups. While the HRT group did experience significantly fewer hip fractures and a lower incidence of heart disease, there was a significant increase in risk for breast cancer, uterine cancer, and blood clots associated with the pHRT group, $X^2(1, n=16608) = 83.6, p<.001$.

Topic: Effects of Leading Questions

Low Substance:

When questioning a person about a past event, the way in which the questions are worded can often affect the way the person remembers the event. Such questions are called "leading questions." Police officers have found that they can influence witnesses' memories for events by asking leading questions. The memory of a witness can actually be changed by a single question. For example, mentioning an item in a question will tend to create a memory of that item, even when no such item existed. Laboratory studies have supported this effect of leading questions on memory. Some such studies have simulated various witness interrogation situations, and have succeeded in replicating the effect of leading questions on witnesses' memories. In these studies, subjects have been successfully caused to recall objects that did not exist or events that never happened.

Consistent High Substance:

When questioning a person about a past event, the way in which the questions are worded can often affect the way the person remembers the event. Such questions are called “leading questions.” In a research study assessing the effect of leading questions on memory, 150 college students were shown a film depicting an automobile accident. In a subsequent questionnaire about the film, subjects were randomly assigned to one of two possible questions: 1) How fast was the car going when it ran the stop sign, or 2) How fast was the car going when it turned right. (There was no actual stop sign shown in the film.) At the end of the questionnaire, all subjects were asked if they had seen a stop sign in the film. Of the subjects asked Question #1, 25 (35%) later reported having seen a stop sign, whereas only 9 (12%) of those receiving question #2 made the same claim, $X^2(1, n = 150) = 4.98, p < .05$

Inconsistent High Substance:

When questioning a person about a past event, the way in which the questions are worded can often affect the way the person remembers the event. Such questions are called “leading questions.” In a research study assessing the effect of leading questions on memory, 150 police officers were shown a film depicting an automobile accident. In a subsequent questionnaire about the film, the officers were randomly assigned to one of two possible questions: 1) How fast was the car going when it ran the stop sign, or 2) How fast was the car going when it sped past the stop sign. At the end of the questionnaire, all subjects were asked if they had seen the car stop in the film. Of the subjects asked Question #1, 25 (35%) later reported having seen the car stop, whereas only 9 (12%) of those receiving question #2 made the same claim, $X^2(1, n = 150) = 4.98, p < .05$

Topic: Effects of Nutrition Education

Low Substance:

Teenagers are known for their poor eating habits. Many high school administrators are aware of this problem and are concerned about their students' eating habits. In an effort to improve their students' eating habits, many high schools have instituted mandatory nutrition classes. To see if the nutrition classes are having any effect, some schools have conducted studies to see if their students' eating habits have changed at all after taking the classes. The studies have investigated if and how the eating habits of the students who have taken the class have changed. The results of these studies have shown that teenagers' eating habits can be improved significantly by taking a nutrition course, and thus it is expected that more school administrators will institute mandatory nutrition classes in their high schools in the future.

Consistent High Substance:

A study was conducted at a high school to examine the effect of a mandatory nutrition class on the students' eating habits. The students in four classes of a Junior level English Composition course were asked to journal their daily eating habits for one week at the beginning of the

semester, and then again at the end of the semester. The comparison groups consisted of those students who were taking the nutrition course concurrently (the treatment group) and those who were not (the control group). The difference between the amount of "junk" food eaten at the beginning of the semester and at the end of the semester was measured and compared between the two groups. Students in the nutrition course significantly reduced their average daily servings of junk food (-2.3) as compared to the control group (-.02), $t(120) = 2.31$, $p < .02$, two-tailed.

Inconsistent High Substance:

A study was conducted at a high school to examine the effect of a mandatory nutrition class on the students' eating habits. The students in four classes of a Junior level English Composition course were asked to journal their daily eating habits for one week at the beginning of the semester, and then again at the end of the semester. The comparison groups consisted of those students who had previously taken the mandatory nutrition course (the treatment group) and those who had not (the control group). The difference between the amount of "junk" food eaten at the beginning of the semester and at the end of the semester was measured and compared between the two groups. Students who had taken the nutrition course significantly reduced their average daily servings of junk food (-0.02) as compared to the control group (-2.3), $t(120) = 2.31$, $p < .02$, two-tailed.

Topic: Social Loafing

Low Substance:

Social loafing is a term used in psychology to describe the fact that individuals will exert less effort on a task when they are working within a group than when they are working alone. For example, if a group of people is pulling on a rope in a tug-of-war, each person will pull with less strength than they would if they were pulling alone. Why would this be? A recent study suggests that it is the fact that the individual's effort is not identifiable when performing in a group, and thus they can get away with exerting less effort. In this study, each person was told that the researchers were able to measure their individual efforts even when performing in a group. Interestingly, when they thought their individual efforts were identifiable, the amount of effort they put forth was exactly the same as when performing alone. Thus, social loafing within in a group can be discouraged if each person is led to believe that his or her individual effort is still identifiable.

Consistent High Substance:

Social loafing is a term used in psychology to describe the fact that individuals will exert less effort on a task when they are working within a group than when they are working alone. In an study of social loafing, researchers hypothesized that the strength of an individual's effort will depend on whether his effort is identifiable or not. In a completely within-subject design, 48 male college students, in groups of six, were asked to shout as loudly as they could. The subjects wore blindfolds and headphones to prevent them from seeing or hearing each other. Subjects

believed they were shouting sometimes alone and sometimes with the group, although they were actually always shouting alone (for ease of measurement). Then subjects were fitted with fake lapel microphones and told that their individual shouting level could then be identified even when shouting with the group. Individuals made 63% as much noise when they believed they were shouting in groups as when shouting alone, $F(1,7) = 55.8$, $p < .0001$, but when they thought they were identifiable within the group, their effort was not significantly different from shouting alone (98%).

Inconsistent High Substance:

Social loafing is a term used in psychology to describe the fact that individuals will exert less effort on a task when they are working within a group than when they are working alone. In a study of social loafing, researchers hypothesized that the strength of an individual's effort will depend on whether his effort is identifiable or not. In a mixed between/within design, 48 male college students, in groups of six, were asked to shout as loudly as they could. The subjects wore blindfolds and headphones to prevent them from seeing or hearing each other. Sometimes they shouted as a group and sometimes alone, but they always thought they were shouting alone. Then subjects were fitted with lapel microphones and told that their individual shouting level could be identified even when shouting with the group. Individuals made 63% as much noise when they were actually shouting in groups as when shouting alone, $F(1,7) = 55.8$, $p < .0001$, but when they thought they weren't identifiable within the group, their effort was not significantly different from shouting alone (98%).

Topic: Effects of Valium on Anxiety

Low Substance:

People who suffer from a phobia typically experience extreme fear or anxiety in the presence of a relatively harmless object or situation. Phobias can be extremely debilitating to the people who suffer from them. While the anti-anxiety drug Valium is typically used on a daily basis to treat generalized anxiety, a daily dose would be impractical for phobic patients, who only experience anxiety in the presence of their phobic object or situation. Therefore, a study was conducted to test the effectiveness of a single dose of Valium on phobic anxiety. The researchers wanted to find out if giving just one dose of Valium would be effective in reducing the anxiety associated with phobias. In the study, a single dose of Valium was given to phobic patients before putting them in situations involving their phobia, such as standing on the observation deck of a tall building. Single doses of Valium were found to be effective in reducing the debilitating anxiety associated with phobias.

Consistent High Substance:

People who suffer from a phobia typically experience extreme fear or anxiety in the presence of a relatively harmless object or situation. Researchers wished to determine whether Valium, an anti-anxiety drug typically used on a daily basis to treat generalized anxiety, would be effective as a single-dose treatment for phobic anxiety. Twenty-four phobic individuals who were not

already taking Valium were recruited to participate in the study. The participants were randomly assigned to a group that received Valium or a group that received a placebo. The participants were then placed in a room with their phobic object and asked to approach the object as closely as possible. The resulting distance of the participant from the object was measured. Participants who were given Valium were able to approach their phobic object to an average distance of 2.3 feet, significantly closer than the subjects in the placebo group at 6.5 feet, $t(22) = 2.825$, $p < .01$, two-tailed.

Inconsistent High Substance:

People who suffer from a phobia typically experience extreme fear or anxiety in the presence of a relatively harmless object or situation. Researchers wished to determine whether Valium, an anti-anxiety drug typically used on a daily basis to treat generalized anxiety, would be effective as a single-dose treatment for phobic anxiety. Twenty-four phobic individuals who were already taking Valium were recruited to participate in the study. The participants were randomly assigned to a group that received Valium or a group that received a placebo. The participants were then placed in a room with their phobic object and asked to approach the object as closely as possible. The resulting distance of the participant from the object was measured. Participants were then given Valium and were able to approach their phobic object to an average distance of 6.1 feet, significantly closer than the subjects who were already taking Valium at 5.2 feet, $t(22) = 1.885$, $p < .10$, two-tailed.